Do performance contracts attract better teachers? A blinded pre-analysis plan

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

September 27, 2018

Contents

1 Introduction

The ability to recruit, elicit effort from, and retain civil servants is a central challenge of state capacity in developing countries. Nowhere is this more evident than in the education sector, where rising access to government schooling has failed to translate into hoped-for learning gains, even as teacher salaries account for the bulk of expenditure on education and a large part of the civil service payroll (Das et al, 2017). Many developing country governments obtain poor skill and effort levels in return for their expenditure on the teaching workforce. For example, the World Bank's Service Delivery Indicators for Uganda suggests that only 20 percent of primary school teachers have mastery of their content, while they are absent from school an average of 27 percent of the time (Wane and Martin 2013). Yet teacher quality has important effects: both immediate, on student learning, and eventual, on later education and labor outcomes (Chetty et al., 2014a,b). Improved skills ultimately affect a country's economic performance (Hanushek and Woessmann, 2012).

Accumulating evidence shows that pay-for-performance contracts can elicit improvements in effort from incumbent teachers, although these results are sensitive to design (Neal, 2011). Individual-teacher performance contracts have had persistent effects in India (Muralidharan, 2012; Muralidharan and Sundararaman, 2011) and Israel (Lavy, 2009) but evidence on school-level incentives is more mixed, e.g., in Kenya (Glewwe et al., 2010). Similarly, evaluations of teacher performance pay in wealthier countries have yielded generally positive results, with incentive effects strongest among relatively inexperienced teachers and diluted as contracts pool teachers (Fryer, 2013; Fryer et al., 2012; Goodman and Turner, 2013; Imberman and Lovenheim, 2015; Lavy, 2009; Sojourner et al., 2014; Springer et al., 2010).

However, little is known about how pay-for-performance contracts affect the *composition* of the civil service.[1] Theory suggests that the extensive-margin effects of performance contracts may be substantial (Lazear, 2003; Rothstein, 2015). This is all the more important given that teacher quality, typically measured as teacher value-added in student learning outcomes, is hard to screen for (Hanushek and Rivkin, 2006; Rockoff et al., 2011). Theoretically, these benefits are balanced by concerns that performance contracts may have signaling (Bénabou and Tirole, 2003) and screening (Delfgaauw and Dur, 2007) effects that crowd out intrinsic motivation, and claims that 'mission matching' may diminish the need for explicit incentives (Besley and Ghatak, 2005).

This project—and the present pre-analysis plan—provides the first prospective experimental evaluation of not only the effort margin but also the compositional effects of pay-for-performance contracts.

A simple framework underpins our intervention and analysis. The production function for student learning depends on, inter alia, two teacher characteristics: skill (pre-determined) and effort (a strategic choice). Teacher utility is decreasing in effort and increasing in both money income and student learning. We refer to the relative importance of student learning in the teacher's utility function as a teacher's intrinsic motivation. Pay-for-performance contracts may then elicit greater

[1] There is a small but growing literature documenting the effects of more general contractual terms (e.g. pay levels) on compositional margins. For instance, in developing countries, higher pay attracts skilled and motivated workers to hardship posts (Dal Bó et al., 2013), and career track opportunities attract competent and motivated community health workers (Ashraf et al., 2016); promises of easy money attract but ultimately disappoint community health workers (Deserranno, 2017). In the U.S., low-performing teachers at risk of firing from a teacher accountability initiative in DC quit on their own (Dee and Wyckoff, 2015) and were replaced by people who did better (Adnot et al., 2016). Meanwhile, there is suggestive evidence both that higher value-added teachers have better earnings opportunities outside the classroom (Chingos and West, 2012) and that their retention is sensitive to pay (Clotfelter et al., 2008). Finally, Figlio and Kenny (2007) report evidence from US counties showing an association between performance pay and student outcomes that comprises both a compositional and incentive margin; Woessmann (2011) provides similar cross-sectional evidence at the country level.

effort from a given teacher, or may change the type of teacher who applies for and is placed in schools, or both.

The intervention takes place in Rwanda's primary education sector. This is an attractive setting for the study, for several reasons. As in most developing countries, teachers represent a large part of the civil service and the bulk of the education budget (Kremer and Holla, 2009). In Rwanda, teacher turnover is high—particularly for primary schools and in rural districts—and replacement of lost teachers is a lengthy process that often leaves schools without qualified replacements for a period of at least a year. Rural districts struggle at times to meet hiring targets, and face acute skill shortages, particularly in the upper-primary grades (Primary 4–6) in which English becomes the official language of instruction. Moreover, there is evidence from the health sector that performance contracts may be effective motivators in the Rwandan context (Basinga et al., 2011) and in fact teachers are currently the only part of the civil service to be exempted from the *imihigo* system that ties a component of pay to measures of performance.

To undertake this study, we worked with the Rwanda Education Board and Ministry of Education to design and implement a pay-for-performance (hereafter P4P) contract based on the 'pay for percentile' scheme (Barlevy and Neal, 2012). Building on extensive consultations and a pilot year, the contract rewards the top 20 percent of teachers using a metric that combines information on both the learning outcomes in teachers' classrooms, what we term *performance*, and three measures of teachers' inputs into the classroom: their *presence* (measured through unannounced visits to schools), their *preparation* (measured through audits of lesson plans), and their *pedagogy* (measured through classroom observations). As described in detail in Appendix B, teachers are ranked on each of these criteria, and their '4P' score is computed as a weighted average, with performance taking half of the weight and presence, preparation and pedagogy weighted equally.

Building on this contract, we undertake a two-tiered experiment (Ashraf et al., 2010; Cohen and Dupas, 2010; Karlan and Zinman, 2009) that randomly assigns labor markets to either P4P or expected-value-equivalent fixed-wage (hereafter FW) *advertisements*, and then uses a surprise re-randomization of *experienced* contracts at the school level to enable estimation of pure compositional effects within each realized contract type. In the first stage—undertaken during recruitment for teacher placements in the 2016 school year—we randomly assigned labor markets to either P4P or FW contracts. We advertised extensively over radio, in flyers, at District Offices, through WhatsApp networks of Teacher Training College alumni, and in job fairs, explaining that all teachers who applied for teaching jobs in the relevant districts and were placed in upper-primary teaching positions would be eligible for the relevant treatments. We then recruited into the study all primary schools that received such a teacher to fill an upper-primary teaching role. In the second stage of our experiment—undertaken once 2016 teacher placements had been finalized—we randomly (re-)assigned these schools in their entirety to either P4P or FW contracts; all teachers, including both newly placed recruits and incumbents, who taught core-curricular classes to upper-primary students would be eligible for the relevant contracts. This was made incentive compatible by effectively buying out recruits' initial offers with a signing bonus, so that no recruit, regardless of her belief about the probability of winning, could be made worse off by the re-randomization.

This pre-analysis plan commits us to a set of primary and secondary hypotheses, as well as to the corresponding variables, statistical specifications, and hypotheses tests. The design of the experiment and our trial registration make clear the core outcomes of the study (see AEARCTR-0002565). In preparing this pre-analysis plan, we go further by making use of a blinded dataset that allows us to learn about a subset of the statistical properties of our data without deriving hypotheses from realized treatment responses, as advocated by, e.g., Olken (2015).[2] This approach achieves

[2]Although we are unable to find examples undertaking such blinding in economics, Humphreys et al. (2013) argue

power gains by choosing from among specifications and test statistics on the basis of simulated power, while protecting against the risk of false positives that could arise if specifications were chosen specifically on the basis of their realized statistical significance in any real-world experiment.[3] For an experimental study in which one important dimension of variation occurs at the labor-market level—and so is potentially limited in power—the gains from these specification choices are particularly important. We demonstrate in this document that, with specifications appropriately chosen, the study design is well powered, such that even null effects would be of both policy and academic interest.

The remainder of this document proceeds as follows. Section 2 sets out the study design in greater detail. Section 3 describes the construction of variables and the blinding procedure, as well as (blinded) summary statistics. Section 4 outlines the primary hypotheses, and the estimation and inference methods used to test them. Section 5 presents key robustness checks. The Appendices provide further information on the motivating theoretical model (Appendix A), the construction of the incentivized performance metric (Appendix B), and the statistical power of the primary hypothesis tests (Appendix C). A second, planned paper focuses on the dynamic consequences of experienced performance contracts for incumbent teachers and teacher retention. The pre-analysis plan for that paper is presented in a separate document.

2 Study Design

The study took place during the actual recruitment of civil service teaching jobs in upper primary in six districts of Rwanda in 2016.[4] The design draws on the 'surprise' two-stage randomizations of Karlan and Zinman (2008), Ashraf, Berry, and Shapiro (2010), and Cohen and Dupas (2010) in credit-market and public-health contexts. Both tiers of this experiment are built around the comparison of two contracts, on top of existing teacher salaries, and managed by Innovations for Poverty Action in coordination with the Rwanda Education Board (REB). The first of these contracts is a pay-for-performance contract, which pays RWF 100,000 (approximately 15 percent of annual salary) to the top 20 percent of upper-primary teachers within a district, as measured by a composite performance metric briefly described in the preceding section and detailed in Appendix B. The second is a fixed wage contract that provides RWF 20,000 to all upper-primary teachers. The second stage (surprise) randomization implies that applicants may experience a different contract from the advertised one to which they applied.[5]

This design gives rise to four distinct types of recruits placed in schools, as summarized in Figure 1. *Potential applicants*—not all of whom are observed—are assigned to either advertised FW or advertised P4P contracts, depending on the labor market in which they reside. Those who actually apply and are placed into schools fall into one of four groups. For example, group 'a' in Figure 1 denotes teachers who applied to jobs advertised as FW, and who were placed in schools assigned to FW contracts, while group 'c' denotes teachers who applied to jobs advertised as FW and were then placed in schools re-randomized to P4P contracts. (As will be discussed in Section

for and undertake a related approach with partial endline data.

[3] The spirit of this approach is similar to recent work Anderson and Magruder (2017) and Fafchamps and Labonne (2017). We give up the opportunity to undertake exploratory analysis in our own dataset, as is possible in their work, because our primary hypotheses are clear from the outset; in return, we avoid having to discard part of our sample, with associated power loss.

[4] "Upper primary" refers to grades 4, 5, and 6 in Rwandan primary schools; the schools themselves typically include grades 1 through 6.

[5] As described in Section 2.2 below, all recruits placed in study schools were offered a retention bonus of RWF 80,000 that ensured that they received *at least* as much under their realized contract as they could have expected under their advertised contract. There were no objections to or refusals of the rerandomziation.

3, using our blinded we know the *sum* of recruits in $a + b$ and the sum of recruits in arms $c + d$, but not the contractual condition under which a recruit applied. For this reason, we refer to these as variables rather than counts of teachers.) The key insight of such a surprise re-randomization is that comparisons between groups a and b, and between groups c and d, allow us to learn about a 'pure' compositional effect of incentive contracts on teacher performance in the school, whereas comparisons along the diagonal of a–d are informative about the total effect of P4P, along both extensive and intensive margins.

Figure 1: Treatment groups among recruits placed in study schools

		Advertised	
		FW	P4P
Experienced	FW	a	b
	P4P	c	d

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

2.1 First-tier randomization: Advertised contracts

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

In January 2016, we discovered that districts actually solicited applications at the slightly coarser district-by-*subject-family* level, aggregating subjects into three subject families that correspond to the degree types issued by Teacher Training Colleges: math and science (TMS); modern

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

languages (TML); and social studies (TSS). We have 18 such labor markets defined by the product of district and subject family. The result of the randomized assignement is that 7 labor markets can be thought of as being in a 'P4P only' advertised treatment (modern language teaching in Gatsibo and Kirehe, math and science teaching in Kayonza and Nygatore, and social studies teaching in Ngoma, Nygatore, and Rwamagama); 7 in a 'FW only' advertised treatment (modern language teaching in Kayonza and Rwamagama, math and science teaching in Kirehe and Ngoma, and social studies teaching in Gatsibo, Kayonza, and Kirehe); and 4 in a 'Mixed' advertised treatment (modern language teaching in Ngoma and Nygatore, and math and science teaching in Gatsibo and Rwamagama). To illustrate this Mixed treatment, an individual living in Ngoma with a qualification to teach modern languages could have applied to the modern languages pool, in which case they would have been eligible for either an advertised post in English on a FW contract, or an advertised post in Kinyarwanda on a P4P contract. In contrast, someone living in Gatsibo with a qualification to teach modern languages would have been subject to the 'P4P only' treatment; he/she could have applied for either an English or Kinyarwanda post, but both would have been on a P4P contract. Empirically, we will consider the Mixed treatment as a separate arm. We will estimate a corresponding advertisement effect but interpret this only as an incidental parameter.

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

2.2 Second-tier randomization: Experienced contracts

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

All upper primary teachers within each school received the new contract. At individual applicant level, this amounted to re-randomization and hence a change to the initial assignment for some new recruits. To ensure that new contracts strictly dominated those advertised at the first tier, all new recruits were told that they would receive a retention bonus of 80,000 RWF if they remained in post during the 2016 school year. Teachers in P4P schools were also told that the 2016 performance award—determined by multiple teacher-input observations as well as beginning- and end-of-year student assessments—was conditional on remaining in post during the 2016 school year, and would be paid early in 2017.

The experiment continued in the same 164 schools for the 2017 school year. Schools were contacted by telephone in February 2017 to remind them of the continuation of the scheme. Teachers in FW schools were told that they would receive the RWF 20,000 award, and teachers in P4P schools were told that the 2017 performance award—calculated in a similar fashion to the 2016 award—would be paid early in 2018. Our enumerators stressed that both payments were conditional on remaining in post during the 2017 school year.

3 Measurement and 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 application decision, to the set (and attributes) of candidates hired into schools, to the learning outcomes that they deliver, and, finally, to their decisions to remain in the schools. In this section, we describe the administrative, survey, and assessment data available for each of these steps in the causal chain. A schematic of these data sources, and their timing in relation to intervention, is laid out in Figure 2.

Our understanding of these data informs our choices of specification for analysis, as will be discussed in Section 4. But crucially, in order to do so without jeopardizing the validity of statistical inference, we have *blinded* the data in a way that makes it impossible for us to mine for results. This scheme has two components, reflecting the need to blind researchers with respect to treatment effects of both the first- and second-stage randomizations. At the student and teacher level, all subjects identifiers (whether of subjects taught in the classroom or of subjects of teacher qualification) are assigned a random number. This was done in a manner consistent across data sources and variables, to preserve patterns of within- and across-subject correlations in outcomes and relative sample sizes. Then, at the school level, post-baseline outcomes are blinded for all schools in the FW arm, effectively by dropping those schools from the post-baseline datasets.[7] This precludes researchers from gaining insight into the treatment effects of the second-stage, school-level randomization.

3.1 Applications

Table 1 summarizes the applications for the newly advertised jobs, submitted in January 2016, across the six districts, using the administrative data described above. Of the 2,185 applications in total, 1,938 come from candidates with a Teacher Training College (TTC) degree—we term such applicants 'qualified', since, at least in principle, such a degree is required for the placements at stake. Candidates were required to sit for a district-specific exam; we denote qualified applicants who do so as having *complete* applications.[8] We present TTC scores, genders, and ages—the other observed CV characteristics—for all qualified applicants in this table (regardless of whether their application includes a district exam score).

Table 1: Application characteristics, by District

	Gatsibo	Kayonza	Kirehe	Ngoma	Nyagatare	Rwamagana
Applicants	390	310	462	381	327	315
Qualified	307	261	458	365	268	279
Qualified DistrictTest	221	177	312	173	162	69
Has TTCscore	294	237	406	338	256	167
Mean TTCscore	0.55	0.55	0.50	0.53	0.54	0.55
SD TTCscore	0.12	0.13	0.19	0.14	0.11	0.10
Complete	221	177	312	173	162	69
Qualified Female	0.54	0.49	0.45	0.52	0.47	0.46
Qualified Age	25.86	26.88	27.24	27.29	26.00	26.66

[7] We drop data from the FW schools because the IPA team required input into the delivery of P4P contracts from the research team.

[8] We observe that a small number of candidates were allowed to sit district exams in spite of not having TTC degrees; we are currently auditing information about the TTC degrees for these individuals, since this should not have been possible in principle.

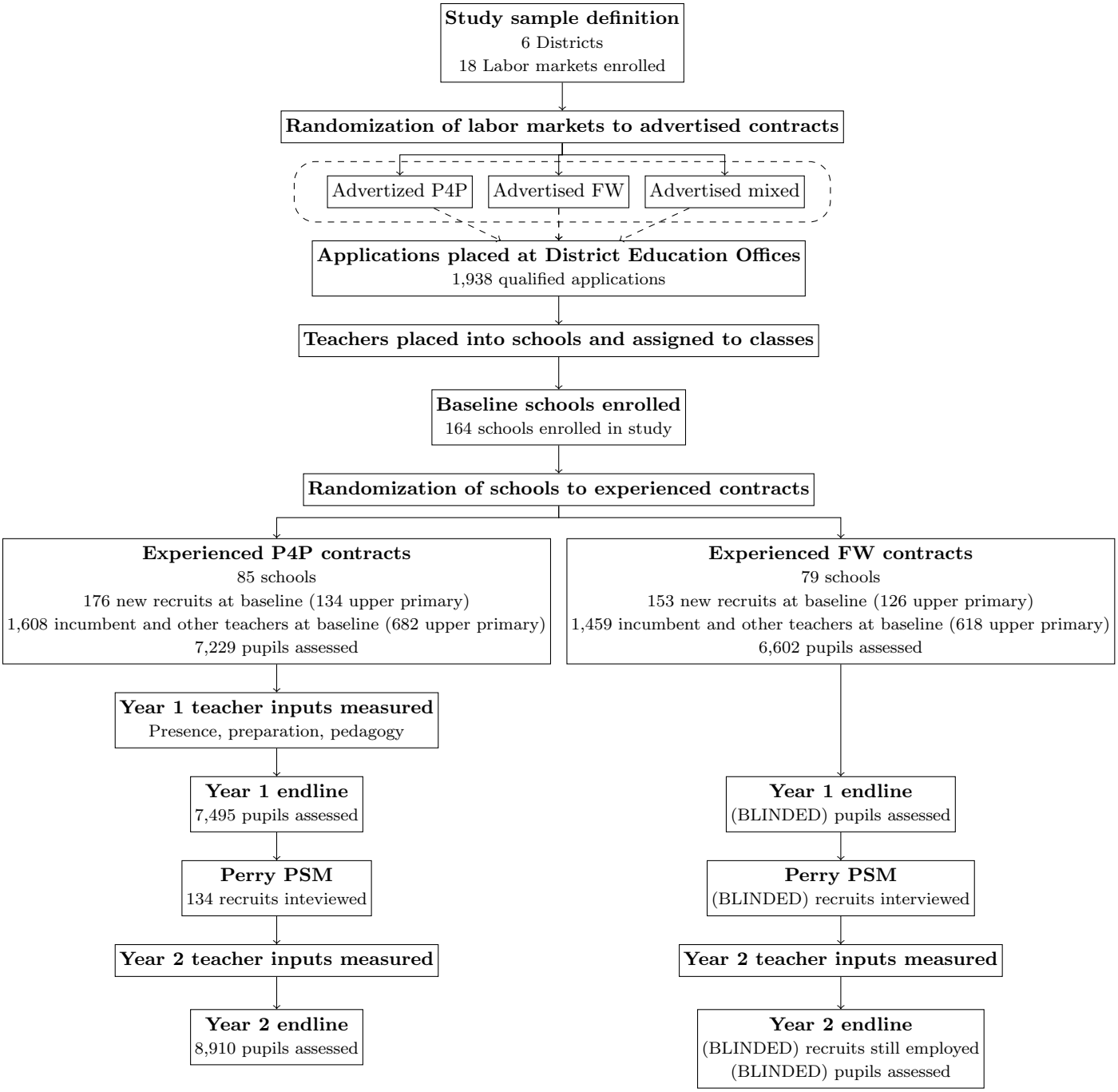


Figure 2: Study profile

The 2,185 applications come from 1,424 unique individuals, of whom 1,194 have a TTC qualification. (Using the blinded data we observe that 465 have blinded qualification type '1'; 344 blinded qualification type '2'; and 385 blinded qualification type '3'). Qualified applicants complete an average of 1.61 applications in study districts, with 62 percent of qualified applicants completing only one application.

In addition to submitting their TTC qualifications, applicants were required to undertake a district-level exam in order to be considered for a post. This final step was added only after applications were submitted, in a change of regulation from the Rwanda Education Board. Each district constructed its own assessment for this purpose, and districts used the same assessment tool across all subjects of application. Consequently, we can use ranks on these assessments relative to ranks in the FW applicant pool for that district as an alternative measure of the relative quality of P4P applicants.

3.2 Teacher attributes

As outlined in the Study Profile (Figure 2), following the application stage, successful applicants to positions in study districts were placed into schools by District Education Officers (DEOs), and were assigned to particular grades, subjects, and streams by the Head Teachers (HTs) in those schools. As described in Section 2, any teacher—whether applicant or incumbent—who was assigned to teach one of the five 'core curricular' subjects in the upper-primary grades of P4, P5, or P6 would be eligible for our intervention. We therefore enrolled all primary schools in study districts in which at least one new recruit had been placed, and where at least one new recruit was assigned to an upper-primary teaching role. We visited these schools at baseline in February 2016, and collected data using three broad types of instruments: school surveys, teacher surveys, and teacher 'lab' measures.

We describe these measures below and in Table 2. In doing so, we separately summarize the attributes of three mutually exclusive types of teachers: recruits, who are formally hired through the district, incumbents, who are teachers that have been in the school in the previous year, and 'other' teachers, who include informal and community hires not affected by the first-stage treatment of P4P or FW advertisements.

School surveys. School surveys were administerd to head teachers, or their deputies, at baseline. These included a variety of data on management practices—not documented here—as well as administrative records of teacher attributes, including ages, genders, and qualifications. These data covered all teachers in the school, regardless of whether they were eligible for the study sample.

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

'Lab-in-the-field' instruments. In addition to measures of teacher characteristics gleaned from surveys, we use a series of incentivized 'lab-in-the-field' tasks to provide additional measures of teachers' preferences for contracts and motivation on the job.

Table 2: Teacher characteristics at baseline

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

Note: For each variable, we report the mean and standard deviation. Numbers of observations reported for each data source.

In a framed version of the *Dictator Game* (Kahneman et al., 1986), each participant was given 2,000 Rwandan Francs (RWF) and was asked how much of this money they wished to allocate towards providing a school supply packet, and how much they wished to keep for themselves. Each student school supply packet was worth 200 RWF, meaning that, in theory, the teacher/head teacher could allocate all 2,000 RWF to providing 10 (randomly chosen) students with a school packet. The purpose of this game was to measure 'other regarding' preferences in a way that would be likely to predict teachers' willingness to allocate their time and effort to student learning. As shown in Table 2, recruits playing this game gave an average of 27 percent of the stake to the schools' students—substantially less than the average donated share of 43 percent by incumbent teachers.

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

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

This grading task was in fact the first stage in a *Competition Game*, based on Niederle and Vesterlund (2007), which has been shown elsewhere to be associated with gender differences in the taste for competitive careers (Buser et al., 2014). Following the fixed-pay grading task that provides our measure of teacher skill, teachers were asked to undertake a second grading exercise. This second grading task took the form of a tournament: only the top 20 percent of teachers within a given subject and district would receive a payout, and the payout would be 5 times the payout for the fixed-pay grading task.[9] Finally, in a third round teachers were allowed to choose between the fixed-pay and tournament payment schemes. Table 2 highlights that nearly two thirds of both recruits and incumbents chose the tournament scheme; this decision (residualized to account for differences in actual ability) will provide a measure of the 'taste for competition' used in the secondary analysis described in Section 5.1.

3.3 Perry public sector motivation

In addition to baseline measures of teacher motivation from the framed dictator game described above, we also used a phone survey to implement the Perry Public Sector Motivation (PSM) instrument (Perry, 1996; Perry and Wise, 1990) for recruits with upper-primary teaching assignments. The Perry PSM instrument has been used in a variety of contexts, and has been found to be predictive of both the compositional response to unconditional salary increases (Dal Bó et al., 2013) and of responses to financial incentives (Callen et al., 2018).

A potentially important difference between this Perry PSM measure and our baseline survey measures is that it was collected during the second year of the experiment, as indicated in Figure 2. Consequently, in this blinded pre-analysis we have access to data only for those recruits in the

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

P4P arm, consistent with the blinding scheme. Moreover, these blinded data do not tell us the advertisement treatment under which the recruits applied. We therefore report means separately for recruits (pooling across the blinded advertisement treatments), incumbents, and other teachers, using data from the P4P arm only.

We administered five of the six subscales of the Perry PSM instrument: 'Commitment to the public interest', 'Social justice', 'Civic duty', 'Compassion', and 'Self-sacrifice'. We do not collect data on the sixth subscale, 'Attraction to policy making', as this was deemed irrelevant for the frontline service-provision roles under study here. On each component of these subscales, respondents' answers are coded on a five-point Likert scale, with 5 representing strong agreement and 1 strong disagreement. We present unweighted subscale means in Table 3.[10] To construct an overall measure of respondents' motivation, we follow Dal Bó et al. (2013) to construct an index of public-service motivation by taking the equally-weighted z-scores of each of the items.[11] This index is presented alongside the subscale means in Table 3. While for purposes of illustration this outcome has been standardized based on the outcomes of all recruits in the realized P4P arm, in our unblinded analysis we will construct this outcome based on the distribution of responses among recruits in the advertised FW arm whose realized contracts were also FW.

Table 3: Perry PSM

	Advertised P4P	Advertised FW
Commitment to public interest	1.79	.
	(0.52)	
Social justice	1.36	.
	(0.42)	
Civic duty	2.17	.
	(0.39)	
Compassion	1.96	.
	(0.44)	
Self-sacrifice	1.74	.
	(0.44)	
PSM index	-0.00	.
	(0.35)	
Observations	134	.

Note: Summary statistics presented here are for all recruits placed in realized P4P schools; outcomes in FW schools are blinded and so not included here.

3.4 Student learning

Student learning was measured via assessments taken at the start and end of the 2016 school year, and the end of the 2017 school year (indexed by {0,1,2}, respectively, in subsequent notation). These student assessments play a dual role in our study: they provide the primary measure of learning for analysis of program impacts, and they were used in the evaluation of teachers in the

[10] Some of the Perry PSM questions are defined so that 'strong agreement' corresponds to lower motivation; in all analyses we reverse scales as necessary so that agreement with the prompt always implies greater motivation.

[11] English-language text of the individual items used and corresponding subscales are provided in Appendix Table D.5, and thd distribution of item-level responses among recruits within the blinded sample are illustrated in Appendix Figure D.6.

realized P4P arm for purposes of performance awards, as discussed in Appendix B.[12]

We developed comprehensive subject- and grade-specific, competency-based assessments for P4, P5, and P6.[13] These assessments were based on the new Rwanda national curriculum and covered the five core subjects: Kinyarwanda, English, Mathematics, Sciences, and Social Studies. We developed one assessment per grade-subject, with students at the beginning of the year being assessed on the prior year's material (and a special P3 assessment developed for the purpose of assessing P4 students at the beginning of the year). Each test aimed to cover the entire curriculum for the corresponding subject and year, with questions becoming progressively more difficult as a student advanced in the test. The questions were a combination of multiple choice and fill-in diagrams.[14]

In each round, we randomly sampled a subset of students from each grade to take the test. These students were sampled from the official school register of enrolled students, which is compiled at the beginning of the year, and which we entered electronically.[15] In Year 1 of the study, both baseline and endline student samples were drawn from the listing at the beginning of the year. This was done to ensure that the sampling protocol did not create incentives for strategic exclusion of students. In Year 2, students were assessed at the end of the year only, and were sampled from the registry as collected in the second trimester.

Student samples were stratified by *streams*; these are subgroups in which they are taught, with students in the same stream staying together for all subjects. In the baseline (Round 0), we sampled a minimum of 5 pupils per stream, and oversampled streams taught in at least one subject by a new recruit to fill available spaces, up to a maximum of 20 pupils per stream and 40 per grade. In rare cases of grades with more than 8 streams, we sampled 5 pupils from all streams. At the Year 1 endline (Round 1), we sampled 10 pupils from each stream: 5 pupils retained from the baseline (if the stream was sampled at baseline) and 5 randomly sampled new pupils. We included the new students to alleviate concerns that teachers in P4P schools might teach (only) to previously sampled students. At Year 2 endline (Round 2), we randomly sampled 10 pupils from each stream from the register for that year.[16] Resulting sample sizes are presented in Table 4.

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.[17] Individual student test results were kept confidential from teachers, parents, head teachers, and Ministry of Education officials, and have only been used for performance award and evaluation purposes in this study.

[12]Because our primary interest in this paper will be the compositional effects of Advertised P4P, and not of Experienced P4P, we are less concerned about the possibility that teaching to this test will bias our results. As a robustness check, our analysis of the impacts of Experienced P4P will use national exam data, which were not part of this incentive metric, but which have the disadvantage of being collected only for P6 students, as a check on our results.

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

[14]In piloting all student tests were administered in English, but we found that P4 students had not yet received the level of English instruction necessary to be adequately measured using an English-based exam. These tests were translated and administered in Kinyarwanda for P4 grades only throughout the study.

[15]The student sample therefore includes some individuals who subsequently left the school.

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

[17]The pre-programmed proctoring videos added additional safeguards to ensure consistency in test administration and timing.

Table 4: Pupil and assessment descriptive statistics

	Round		
	Baseline	Round 1	Round 2
Schools	85	85	85
Streams	840	.	924
Total upper-primary pupils	34,976	.	38,322
Pupils sampled for test	7,591	8,296	9,216
Pupils taking exam	7,229	7,495	8,910
Student-subjects	36,139	37,475	44,550
E[z]	0.01	0.00	0.00
Var[z]	0.83	0.78	0.79

Note: Descriptive statistics for P4P schools only. Enrollment figures taken from official pupil registration data, updated annually, and hence not collected at the year 1 endline.

Responses were used to estimate a measure of student learning (for a given student in a given round and given subject in a given grade) based on a *two-parameter Item Response Theory (IRT) model*, which was estimated using Stata's `irt 2pl` command. In the blinded data, these are estimated using the full sample of pupils at baseline, and the sample of pupils in the P4P arm only at Year 1 and 2 endlines, but for purposes of precision the IRT model will be re-estimated using the full Year 1 and 2 samples in the final analysis. We use empirical Bayes estimates of student ability from this model as our measure of a student's learning level in a particular grade.

3.5 Teacher inputs

We collected data on teachers' inputs into the classroom. This was undertaken in P4P schools only during Year 1, and in both P4P and FW schools in Year 2. These measures contribute to the incentivized teacher performance metric in P4P schools, as described in Appendix B.

Our composite teacher performance metric is based on three input measures (teacher presence, lesson preparation and pedagogical practice), and one output measure (student performance)—the '4Ps'. Here we describe the input components measured.

To assess the three inputs, P4P schools received three unannounced surprise visits: two spot checks during Summer 2016, and one spot check in Summer 2017. During these visits, Sector Education Officers (SEOs) from the District Education Offices (in Year 1) or IPA staff (for logistical reasons, in Year 2) observed teachers and monitored their presence, preparation and pedagogy with the aid of specially designed tools.[18] FW schools also received an unannounced visit in Year 2, at the same time as the P4P schools. Since all post-treatment measures are blinded for FW schools, our blinded pre-analysis dataset includes the outcomes of these visits for P4P schools only. We describe them in Table 5.

Presence is defined as the fraction of spot-check days that the teacher is present at the start of the school day. The SEOs recorded teacher presence after speaking with the head teacher at the

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

Table 5: Measures of teacher inputs in P4P schools

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

Notes: Descriptive statistics presented are for upper-primary teachers in P4P schools only. Outcomes in FW schools are blinded. Overall score for classroom observation is average of four components: Lesson objective, Teaching activities, Use of assessment, and Student engagement, with each component scored on a scale from zero to three.

start of the school day during each unannounced visit. In order for the SEO to record a teacher present, the head teacher had to physically show the SEO that the teacher was in school rather than relying on an attendance roster.

Lesson *preparation* is defined as the planning involved with daily lessons, and is measured through a review of teacher written weekly lesson plans. Prior to any spot checks, teachers in grades P4, P5, and P6 in P4P schools were shown how to fill out a lesson plan in accordance with REB guidance.[19] Specifically, SEOs visited schools and provided teachers with a template to help prepare three key components of a lesson—write out the lesson objective, list the instructional activities, and list the types of assessment that will be carried out—and then, in a 'hands-on' session, enabled teachers to practice writing lesson plans using this template before incorporating it in their daily teaching practice. During the SEO's unannounced visit, he/she then collected the daily lesson plans (if any had been prepared) from each teacher. Field staff subsequently used a lesson planning scoring rubric to provide a subjective measure of quality. Because a substantial share of upper-primary teachers do not have a lesson plan on a randomly chosen audit day, we use the presence of such a lesson plan as a summary measure in both the incentivized contracts and as an outcome for anlaysis.

Pedagogy is defined as the practices and methods that teachers use in order to impact student learning. We collaborated with both the Ministry of Education and REB in May and June 2015 to develop a monitoring instrument to measure teacher pedagogy through classroom observation. Our classroom observation instrument measured objective teacher actions and skills as an input into scoring teachers' pedagogical performance, using a rubric adapted from the Danielson Framework for Teaching, which is widely used in the U.S. (Danielson, 2007). The observer evaluated the teachers' effective use of 21 different activities over the course of a full 45-minute lesson.[20] Based on these observations and a detailed rubric, the observer provided a subjective score, on a scale representing mastery from zero to three, of four components of the lesson: communication of lesson objectives, delivery of material, use of assessment, and student engagement.[21] The teacher's incentivized score, as well as our measure of pedagogy, is defined as the average of these ratings across the four domains.

3.6 Job fairs

Data collection at job fairs. Although not part of our core analysis, we will also report results using data collected at our TTC job fairs. During each of the three job fairs that were held in December 2015—during the application period affected by the intervention—we invited attendees to participate in a survey and to play the behavioral games. The survey and games were identical to those used with teachers at baseline.

Participants in the job fairs provide a pool of *potential* applicants to FW and P4P positions. This is useful because in general the full pool of potential applicants for each position is not observed. The downside is of course that job-fair participants have 'selected in' to attend this informational setting and so are not necessarily representative of the full pool. Attributes measured at job fairs can be linked through administrative data to subsequent application decisions.

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

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

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

Table 6: Job-fair participants

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

4 Empirical specifications and inference

We set out to test six distinct questions. Each of these has a corresponding primary hypothesis, and a small number of associated secondary hypotheses that represent alternative measures or mechanisms. These hypotheses are:

I. Advertised P4P induces differential application qualities;

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

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

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

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

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

In this section, for each of these hypotheses (and their corresponding primary or secondary implementations), we address four questions: (a) What outcome measure will be used? (b) On what sample will this be estimated? (c) What test statistic will be used? And (d) How will inference be undertaken on this test statistic? We summarize these design features in Table 7 and provide details of each below.

Using only the first-tier randomization in advertised contracts to look for direct evidence of recruitment effects, our experimental design permits the analysis of applicant characteristics irrespective of placement outcomes (Hypothesis I) as well as a more detailed analysis of the *pre-job* characteristics of those placed in schools—specifically, measures of skills (Hypothesis II) and motivation (Hypothesis III). A central contribution of this paper is that our empirical strategy does

Table 7: Summary of hypotheses, outcomes, samples, and specifications

Outcome	Sample	Test statistic	Randomization inference
Hypothesis I: Advertised P4P induces differential application qualities			
*TTC exam scores	Universe of applications	KS test of eq. (1)	$\mathcal{T}^A$
District exam scores	Universe of applications	KS test of eq. (1)	$\mathcal{T}^A$
TTC exam scores	Universe of applications	t_A in eq. (2)	$\mathcal{T}^A$
TTC exam scores	Applicants in the top $\hat{H}$ number of applicants, where $\hat{H}$ is the predicted number of hires based on subject and district, estimated off of FW applicant pools	t_A in eq. (2)	$\mathcal{T}^A$
TTC exam scores	Universe of application, weighted by probability of placement	t_A in eq. (2)	$\mathcal{T}^A$
Number of applicants	Universe of applications	t_A in eq. (3)	$\mathcal{T}^A$
Hypothesis II: Advertised P4P affects the observable skills of placed recruits in schools			
*Teacher skills assessment IRT model EB score	Placed recruits	t_A in eq. (4)	$\mathcal{T}^A$
Hypothesis III: Advertised P4P induces differentially 'intrinsically' motivated recruits to be placed in schools			
*Dictator-game donations	Placed recruits	t_A in eq. (4)	$\mathcal{T}^A$
Perry PSM instrument	Placed recruits retained through Year 2	t_A in eq. (4)	$\mathcal{T}^A$
Hypothesis IV: Advertised P4P induces the selection of higher-(or lower-) value-added teachers			
*Student assessments (IRT EB predictions)	Pooled Year 1 & Year 2 students	t_A in eq. (6)	$\mathcal{T}^A$
Student assessments	Pooled Year 1 & Year 2 students	t_A and t_{A+AE};	$\mathcal{T}^A$
		t_{AE} in eq. (7)	$\mathcal{T}^A \times \mathcal{T}^E$
Student assessments	Year 1 students	t_A in eq. (6)	$\mathcal{T}^A$
Student assessments	Year 2 students	t_A in eq. (6)	$\mathcal{T}^A$
Hypothesis V: Experienced P4P creates incentives which contribute to higher (or lower) teacher value-added			
*Student assessments (IRT EB predictions)	Pooled Year 1 & Year 2 students	t_E in eq. (6)	T^E
Student assessments	Pooled Year 1 & Year 2 students	t_E and t_{E+AE};	$\mathcal{T}^E$
		t_{AE} in eq. (7)	$\mathcal{T}^A \times \mathcal{T}^E$
Student assessments	Year 1 students	t_E in eq. (6)	$\mathcal{T}^E$
Student assessments	Year 2 students	t_E in eq. (6)	$\mathcal{T}^E$

Continues...

Table 7, continued

Outcome	Sample	Test statistic	Randomization inference
Hypothesis VI: Selection and incentive effects are apparent in the 4P performance metric			
*Composite 4P metric	Teachers, pooled Year 1 (experienced P4P only) & Year 2	t_A in eq. (8)	$\mathcal{T}^A$
Composite 4P metric	Teachers, pooled Year 1 (experienced P4P only) & Year 2	t_A and t_{A+AE}; t_E and t_{E+AE}; t_{AE} in eq. (9)	$\mathcal{T}^A$ $\mathcal{T}^E$ $\mathcal{T}^A \times \mathcal{T}^E$
Barlevy-Neal rank	As above		
Teacher attendance	As above		
Classroom observation	As above		
Lesson plan (indicator)	As above		

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

not hinge only on the ability to measure teacher quality ex ante, which is important given the widely recognized challenge of finding good pre-job teacher characteristics to predict on-the-job performance. Combining the first and second-stage randomization we can look at teachers' on-the-job performance, holding constant the contractual environment into which they were placed to test for pure compositional effects on teacher value-added (Hypothesis IV). Using the second-stage randomization, we can also test for incentive effects of the experienced contracts on subsequent performance (Hypothesis V). Finally, using both the first and second-stage randomization, we can also test whether these selection and incentive effects are present in the composite '4P' performance metric (Hypothesis VI).

In general, the approach to null hypothesis significance testing deployed here will use Fisherian randomization inference (see, e.g., Gerber and Green, 2012; Imbens and Rubin, 2015, for overviews). We will report significance levels for the test statistics in question, and, where the relevant test statistics are regression parameters, we will present 95 percent confidence intervals derived by inverting the Fisher randomization test.

The rationale for this approach is (at least) threefold. First, it provides a unifying framework for hypothesis testing that is exact and avoids reliance on asymptotic results, when clusters (defined as units of randomization) are sometimes small and outcomes are both non-normal and correlated within clusters. Randomization inference takes into account the extent of interdependence by construction. Second, it allows us to focus on the specific source of uncertainty that relates to the hypothesis being tested. For example, when testing hypotheses about pure compositional effects, we will hold fixed the second-stage (experienced) P4P treatment and permute only the advertised, recruitment treatment. If the true effect of the experienced treatment is non-zero, this avoids injecting statistical uncertainty that is not due to the hypothesis being tested, and so helps to improve power and ensure appropriate coverage of confidence intervals (DiCiccio and Romano, 2017). (The final column of Table 7 indicates the specific hypothesis tests where this one-dimensional permutation will be employed.) And third, to the extent that our secondary hypotheses involve multiple tests, randomization inference provides a natural approach to testing joint significance. For secondary outcomes or specifications within a given hypothesis, in addition to reporting tests based on these coefficients, we will test the hypothesis that all secondary test statistics within that hypothesis are jointly equal to zero, using the minimum p-value across outcomes as a test statistic (Young, 2017).

4.1 Hypothesis I: Advertised P4P induces differential application qualities

A central question for the study is whether (advertised) performance contracts attract "better" teachers as applicants to jobs. Theory is ambiguous about whether P4P contracts will induce a greater *number* of applications in absolute terms. Neither does it imply that applications under one contract should strictly dominate applications on the other. In general, the volume and relative aptitude of application pools under alternative contracts will depend on the interplay of potential teachers' beliefs about their own abilities and their outside opportunities (as well as hiring rules, if application costs are important). Consequently, in this domain, our primary hypothesis is a test for equality in distributions across the application pools induced by the experimental treatments.

The consequences of any induced change in the distribution of applicant qualities will depend on the selection rule used by government, which adds a further element of context-specificity to the results.[22] Even in Rwanda, in the long run the government might find it optimal to respond

[22]We thank Fred Finan for this point. Note also that if the selection rule responded to the treatment, then the downstream results of the following subsections would no longer be interpretable as a pure labor-supply effect, but would also encompass government's hiring response.

to changes in labor supply by altering the mapping from formal qualifications to placement outcomes. This further increases our interest in testing for distributional differences across advertised treatment arms that capture changes in the pattern of labor supply, without overlaying on these a specific (demand-side) mapping.

We have two proxies for teacher ability among the universe of applicants to districts and subjects covered by the experiment: *official TTC leaving exam scores* and *district applicant exam scores*. District applicant exams were hastily developed for the first time in response to a Rwanda Education Board edict in January 2016, differ across districts, and do not test domain-specific knowledge. Given concerns about their quality as predictors of teacher ability and consistency across districts, we take the TTC leaving exam score among applicants as primary, and relegate analysis of district applicant exams to secondary standing.

Our proposal is to compare distributions of the first of these proxies for teacher ability across subgroups of applicants defined by the first-tier randomization (advertised contracts) using the non-parametric Kolmogorov-Smirnov test.[23] Denote F_{FW} and F_{P4P} as the empirical distribution functions of TTC leaving exam scores for individuals who applied for new posts that were advertised under the FW and P4P protocols respectively. The prediction from the theory in Appendix A is that these two distributions should differ, and that F_{P4P} should first order stochastically dominate F_{FW}.

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

where $\hat{F}_W(y)$ is the empirical cumulative distribution function of applicant characteristic y in treatment arm W, evaluated at some specific value y.

As noted in our general approach to inference above, we will test the statistical significance of this difference in distributions by randomization inference. To do so, we repeatedly sample from the set of potential[24] advertised P4P treatment assignments $\mathcal{T}^A_{qd}$. For each such permutation, we calculate the Komologorov-Smirnov test statistic T^{KS}. The p value for test statistic T^{KS} is then given by the share of such test statistics larger in absolute value than the test statistic estimated from the actual assignment.

To the extent that differences in the applicant pool are induced by the contract type, government-side selection rules may have consequences for the attributes of hired teachers. To study consequences under alternative selection rules, we estimate a series of *weighted* regressions of the form

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

where y_{iqd} denotes the TTC exam score of applicant teacher i with qualification q and district d. Treatment T^A_{qd} denotes the contractual condition under which a candidate applied.[25]

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

[23] As noted by Imbens and Rubin, "because [the KS statistic T^{KS}] is a scalar function of the vector of assignments and the vector of observed outcomes, it is a valid test statistic [for use in randomization inference]." (Imbens and Rubin, 2015, p. 70).

[24] Assignments had to satisfy criteria regarding the extent of imbalance across districts and subjects that they would induce, so not all possible simple randomization of district-subjects were admissible. The selected advertised P4P assignment vector was chosen from a large number of admissible randomizations by simple random sampling.

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

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

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

For each of these secondary hypotheses, the weighted regression parameter τ_A estimates the differences in (weighted) mean applicant skill induced by advertised P4P. To undertake inference about this difference in means, we use randomization inference, sampling repeatedly from the set of *potential* assignments of advertised P4P, $\mathcal{T}^{\mathcal{A}}$. Following Chung and Romano (2013), we studentize this parameter by dividing it by its standard error to control the asymptotic rejection probability against the null hypothesis of equality of means. These tests will be two-sided tests.[26]

Finally, we test for differences in the number of applicants by treatment status, conditional on district and subject-family fixed indicators. We do so with a specification of the form

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

where q indexes subject families and d indexes districts; N_{qd} measures the number of qualified applicants in each district.[27] As above, we obtain the studentized test statistic t_A by dividing the estimated coefficient τ_A by the analytical estimate of its standard error, and use this t-statistic in our randomization inference.

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

As described in Section 3, our primary measure of teacher skills is the Grading Task administered to all placed recruit and incumbent teachers whose assignments at baseline included at least one upper-primary subject. In particular, we use Empirical Bayes predictions from an IRT model of teacher ability in each subject,[28] which we denote by z_{iqd} for teacher i with qualification q and district d. We then estimate impacts on average recruit skill levels using a regression of the form

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

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

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

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

where γ_q and δ_d again denote coefficients on a vector of subject and district indicators, respectively. Inference for the key parameter, τ_A, is undertaken by performing randomization inference for alternative assignments of the recruitment treatment, T^A_{qd}. Following Chung and Romano (2013) and as above, we studentize the regression coefficient of interest τ_A and take the absolute value of this t-statistic, t_A, as the test statistic in our randomization inference procedure.

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

In addition to the possibility that P4P contracts may select teachers on the basis of skill, such contracts may also change the distribution of intrinsic motivation among the pool of applicants. To test for such effects, we use a specification analogous to equation (4), again, conducting inference for the sharp null of no effect using randomization inference on the recruitment treatment, T^A. Specifically, for a measure of teacher preferences x_{iqd}, we estimate

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

As above, we use the studentized t-statistic, $t_A = \tau_A / s_A$, where s_A is the estimated standard error, as the test statistic of interest in our randomization inference procedure.

Our primary measure of intrinsic motivation is the share of the stake allocated by a teacher to the school in the baseline Dictator Game described in Section 3. In addition, we consider the Perry Public Sector Motivation (PSM) score as a secondary measure of the teacher's intrinsic motivation. Though the PSM score has the advantage of comparability to other papers in the literature (e.g., Dal Bó et al., 2013), we treat this with caution given that it was collected only in the second year of the study.

4.4 Hypothesis IV: Advertised P4P induces the selection of higher-(or lower-) performing teachers, as measured by the learning outcomes of their students

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

The primary measure of student learning which we deploy in this section is the Empirical Bayes prediction of student ability, based on the IRT model of student assessments. This is observed at the student-subject level; each sampled student takes an assessment in all five core subjects. We denote this measure by z_{jbkgsr}, for student j in subject b, stream k, grade g, school s, and round r from which student j's outcome is drawn. (Since streams are nested within grades, we suppress an index for grades to reduce notation.) The advertized P4P treatments about which a given student's performance is informative depend on the identity of the teacher teaching her particular subject, and the qualification type and district of that teacher; we denote this by T^A_{qd} for teacher i with qualification type q in district d, and suppress the dependence of teacher's qualification q on the subject b, stream k, school s, and round r, which implies that $q = q(bksr)$. Experienced treatments are assigned at the school level, and denoted by T^E_s.

Since recruited teachers are observed for two years after their placement in schools, our primary estimand is the impact of their recruitment treatment on the placed teacher's annual value added,

holding constant the realized contractual treatment of the school into which they were placed. We will therefore pool data across the two rounds and estimate as primary a specification of the type

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

where δ_d denotes the coefficient on a vector of district indicators. The variable I_i is an indicator for whether the teacher i was an incumbent in the year of experimental recruitment.[29] The ANCOVA control $\bar{z}_{ks,r-1}$ is a vector of mean lagged assessment outcomes for students in that stream; we discuss this in greater detail below, but note here that the parameter ρ_{bgr} differs by grade-subject, allowing the predictive value of, say, past performance in mathematics to be different when predicting current mathematics performance than when predicting current performance in social studies.[30] This specification seeks to maximize power for the ability to reject the null of no recruitment effects by pooling recruits placed in experienced P4P and experienced FW treatments.

However, the theoretical framework set out in Appendix A makes clear that the impact of the recruitment treatment on realized value-added should depend on the contractual environment into which teachers are placed.[31] Consequently, we also estimate a secondary specification that allows recruitment effects on teacher value-added to differ by experienced treatment. This model takes the form

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

Here, the compositional effect of advertised P4P among recruits placed in FW schools is given by τ_A (a comparison of on-the-job performance across groups a and b, as defined in Figure 1). Likewise, the compositional effect of advertised P4P among recruits placed in P4P schools is given by $\tau_A + \tau_{AE}$ (a comparison of groups c and d).

Students of incumbents (and other non-recruits) contribute to the power of the primary and secondary specifications in two ways. Most obviously, since the estimated parameters ρ_{br} and δ_d are pooled across classes taught by recruits and incumbents, inclusion of incumbents will contribute to the precision of those parameters if the pooled model is correctly specified. But more fundamentally, the blinded simulations reported in Appendix C demonstrate that there are non-negligible power gains to be had from estimating a model that incorporates students in both incumbents' and recruits' classrooms. Intuitively, in these specifications, outcomes of incumbents' students are informative because they provide information about school-wide shocks to learning.

Simulations in Appendix C consider the power of several classes of models for estimating equations (6) and (7): OLS, random effects, fixed effects, and linear mixed effects models (the last of these assume normally distributed disturbances, including random effects at various levels, and estimate parameters by maximum likelihood). Of these, the linear mixed effects model with random effects at the student-round level provides the greatest power: for example, the standard deviation of the pooled estimate for τ_A is a mere 0.025 standard deviations under the 'sharp' null of no treatment effects used for illustration. Further, the coefficient on the *experienced* P4P remains identified

[29] In specifications that include new teachers who were placed in 2017—after the recruitment treatment finished—these will be pooled with the incumbents since they were not exposed to experimental contractual offers; we also pool non-recruits who are new to a school but were not subject to the program's incentives at the application stage (e.g., transfers and temporary teachers) with incumbents.

[30] Given that the timing between each round's 'baseline' and 'outcome' tests differs, we also allow these coefficients to differ by round of the study, r.

[31] Compare equations (11) and (12) in Appendix A: the compositional effect of advertised P4P among recruits placed in FW schools is $e^F(\tau^P)\theta^P - e^F(\tau^F)\theta^F$, whereas the compositional effect among recruits placed in P4P schools is $e^P(\tau^P, \theta^P)\theta^P - e^P(\tau^F, \theta^F)\theta^F$. Intuitively, the skill margin has more of an impact under experienced P4P than experienced FW.

in this model—by contrast with models that have fixed effects at levels of the school or below—and that coefficient, too, is estimated relatively precisely by the linear mixed effects (henceforth LME) model. We therefore use this LME estimator, with random effects at the student-year level and district fixed effects, to estimate parameters in equations (6) and (7).

In both the pooled and differential effects models of equations (6) and (7), we include stream-mean lagged test scores, $\bar{z}_{ks,r-1}$, as an ANCOVA control. When the outcome is year-one endline assessments, $r = 1$, these lagged assessment scores correspond to baseline scores, but when the outcome is year-two assessments, lagged test scores correspond to year-one outcomes.[32] We use stream k mean lagged assessment scores, rather than lags specific to student j, because the rotating sample of students within a given stream means that a restriction to students for whom consecutive test scores are available would have a restrictive effect on sample size, and power, though we can do so as a robustness check.

We include this ANCOVA control for several reasons. First, our estimand is teacher quality, as measured by the average annual value added of placed recruits, rather than the cumulative effect of recruitment. If we were *not* to include controls for once-lagged test scores, the estimated recruitment effect, τ_A, would represent a strange weighted combination of the one-year impact of being taught by a recruit and the cumulative effect of being taught by a combination of recruits and incumbents. Thus, for example, de Ree et al. (2018) estimate a specification that includes indicators for all possible treatment exposures over a two year period, to parse out these combinations, but they note that these exposure paths are not randomly assigned. Second, assignment of students to teachers is potentially non-random with regard to their recruitment treatment. To the extent that this is true, failure to accurately control for students' incoming test scores can bias estimates of annual value added, as is well understood in the value added literature (Kane and Staiger, 2008; Rothstein, 2009, 2010). Thus, to proxy as fully as possible, given our data, for potential non-random sorting of recruits into subject-streams of differential quality, we control for the full vector of lagged assessment outcomes. Simulations undertaken in Appendix C confirm that our model returns unbiased estimates of a non-zero coefficient, τ_A, when this is imposed as the data generating process.

As with other regression-based tests, we studentize the relevant regression coefficients by dividing them by their estimated standard errors when conducting randomization inference Chung and Romano (2013). To be precise, because the LME model is estimated by maximum likelihood, the resulting test statistic is a z statistic rather than a t statistic. The fact that this statistic is not actually compared to that reference distribution for purposes of inference makes this distinction less consequential here than in a case where, say, analytical standard errors were used. To avoid confusion in light of the fact that we have denoted student test scores by z, we refer to these standardized coefficients by t in Table 7.

In conducting randomization inference, we permute only those dimensions of treatment that identify that coefficient (DiCiccio and Romano, 2017). Thus in tests of hypotheses involving only τ_A, in the pooled specification, we conduct randomization inference with respect to the advertised P4P treatment $T^A \in \mathcal{T}^A$.[33] This approach will also be used when testing the existence of a compositional effect in the interacted regression of equation (7) in either the experienced FW arm (parameter τ_A) or the experienced P4P arm (parameter $\tau_A + \tau_E$). To test whether the effect of advertised P4P is

[32] A similar issue is confronted in other studies in which experimental treatments impact a subset of teachers in schools, with students potentially endogenously sorted into the classrooms of the affected teachers; see, e.g., de Ree et al. (2018).

[33] In the presence of a true treatment effect of experienced P4P, τ_E, permuting only T^A improves power by reducing noise introduced from the permutation of treatment T^E. Noise arising from imposing a *false* null that $\tau_E = 0$ is incidental to the hypothesis that $\tau_A = 0$.

statistically significantly different under experienced P4P versus FW, we undertake randomization inference with regard to both stages of treatment, so that we draw pairs of vectors T^A, T^E from the set $\{\mathcal{T}^A \times \mathcal{T}^E\}$ of feasible randomizations across both tiers of the experiment.

4.5 Hypothesis V: Experienced P4P creates incentives which contribute to higher-(or lower-)performing teachers, as measured by the learning outcomes of their students

In addition to the compositional effect of advertised P4P, we are interested in the incentive effect of assignment to a P4P school upon recruits.[34] Our primary test for these impacts uses the pooled specification of equation (6), where the incentive effect of experienced P4P among (all) recruits is captured by the parameter τ_E. As above, we use the associated studentized coefficient t_E (technically, a z statistic given the LME model is estimated by maximimum likelihood), as the test statistic in randomization inference.

We consider two secondary variants of this hypothesis. First, as the theoretical model in Appendix A makes clear that the impact of the experienced treatment on realized value-added should depend on the contractual environment under which teachers were recruited.[35] We therefore use the interacted specification of equation (7) to separately estimate and test for responses to experienced P4P among the cohorts recruited under advertised FW and under advertised P4P, and to test for the difference between them. Second, we use the pooled specification to estimate impacts separately by year of exposure. Randomization inference with respect to the set of possible assignments of $T^E \in \mathcal{T}^E$ remains the basis for inference (except when we test for a statistically significant difference in incentive effects by advertised treatment).

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

As outlined in Appendix A, theory predicts that potential teachers should be attracted to contracts on which they expect to perform well. If the composite '4P' performance metric is not perfectly aligned with the production of learning outcomes, it may be the case that impacts on students' assessed learning provide the leading policy-relevant estimand but are not the most direct test of the theory. In order to test the theory more directly, as well as to observe teachers' proximate responses to contracts, we propose to estimate the compositional effect (Hypothesis IV) and incentive effect (Hypothesis V) using the composite 4P performance metric as the estimand. Details of the construction of this composite metric can be found in Appendix B.

The sample for these estimates is complicated by the fact that we only observe teacher inputs in *experienced* P4P schools in Year 1 of the study. Still, there is variation in the *advertised* contractual treatments of recruits observed in Year 1, since both types are placed in P4P schools. Thus, to improve power, our primary tests will pool data across Year 1 and Year 2. As a robustness check, will also estimate this specification (without the round indicator) on Year 2 data only.

Our interest in this analysis lies in the compositional effect of P4P. Mirroring the specifications above, we will estimate a regression of the form

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

[34] Impacts of experienced P4P among incumbents are an object of study in a separate paper.

[35] Compare equations (13) and (14) in Appendix A: the incentive effect of experienced P4P among recruits selected under FW is $e^P(\tau^F, \theta^F)\theta^F - e^F(\tau^F)\theta^F$, whereas the incentive effect among recruits selected under P4P schools is $e^P(\tau^P, \theta^P)\theta^P - e^F(\tau^P)\theta^P$.

where m_{iqsdr} is the composite performance metric for teacher i, with qualification q, in school s of district d, as observed in post-treatment round $r \in \{1, 2\}$. Parameters γ_q, δ_d, and ψ_r denote coefficients on subject-of-qualification,[36] district, and round indicators. We will use the pooled specification of equation (8) as the primary test for the existence of compositional effects, where the pooled coefficient τ_A provides our most powerful available test for violations of the (sharp) null hypothesis of no compositional effect, if the truth is that there are (similarly signed) compositional effects in both experienced P4P arms.

To allow compositional differences to depend on the experienced treatment, we will then estimate a secondary specification that separates out the compositional effects by the experienced P4P treatment

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

By contrast with the models used to study the same parameters when looking at impacts on student learning (Hypotheses IV and V), a linear mixed effects model with student-level random effects is no longer applicable: outcomes are constructed at the teacher level, and given their rank-based construction, normality does not seem a helpful approximation to the distribution of error terms. We therefore estimate equations (8) and (9) with a round-school random-effects estimator to improve efficiency; this model had performed comparatively on test-score data. However, the inferential strategies—and in particular the particular permutations of treatments used for inferential purposes on each parameter—will mirror those in Hypothesis IV and V. Finally, as further secondary specifications we will repeat the analysis for each of the four components of the performance metric separately.

5 Robustness checks and supplementary analyses

Here, we describe additional analyses that inform interpretation and robustness of our primary hypotheses.

5.1 Broader determinants of potential teachers' labor supply decisions

Our basic theoretical model (as sketched out in Appendix A) focuses on labor supply decisions of a risk-neutral individual with full knowledge of their (ability, motivation) type and preferences defined only over effort and income. In reality, of course, there may be many other factors that determine the value of a potential contract to an individual.

These other determinants of contractual choice are potentially important to our experiment for several reasons. Policy-makers may value these attributes directly: for example, they may value the equal representation of each gender in the teaching workforce, or they may particularly value younger teachers, etc. Further, these characteristics may contribute to the effectiveness of a teacher—whether unconditionally or in response to a specific contract assignment—such that selection on these attributes has direct consequences for the learning outcomes that any given contract produces. Even if such attributes affect only the taste for a given contract, but not teacher effectiveness per se, insofar as they may be *correlated* with other characteristics that affect value added, they can alter and even overturn theoretical predictions about compositional effects from a simpler, more tractable, theoretical model.

As we have no a priori way to sign such correlations, we collected data on a range of characteristics that we judged to have the potential either to be consequential for selection into contracts,

[36] Qualification indicators will be set to zero for incumbent teachers.

and—potentially but not necessarily—to be directly consequential for learning outcomes. Unless otherwise noted, these measures were collected in the baseline survey administered to recruits placed into schools, and they are described in greater detail in Section 3.2 and Table 2.

For each of the following attributes, we will test for mean differences in these characteristics among the placed recruit sample:

- Gender: an indicator for whether the placed recruit is female;
- Age: in years;
- Risk aversion: cardinal index of lottery chosen in the lottery-choice game, from 1–5, with 5 representing the riskiest option;
- Taste for competition: an indicator for whether the respondent chose the tournament (rather than the fixed-wage) payout structure in the Competition Game;
- Locus of control (Rotter index);
- Self-esteem (Rosenberg index);
- Big five index.

To do so, we will take these attributes as outcomes in the specification in equation (4), and use the associated inferential strategy of Hypotheses II and III. To the extent that any of these factors are selected for differentially by the advertised treatments, we will assess whether they are associated with measures of motivation or skill in the sample of hired recruits. Further, we will assess whether these factors predict teacher performance, either unconditionally or in terms of teachers' responsiveness to experienced contracts.

5.2 Demand-side mechanisms

The random assignment of labor markets to advertised contracts allows for a clean experimental test of the null hypothesis that there is no compositional effect on the applicant pool that results. However, all effects further down the causal chain are conditioned on the responses of the government to these applications. Consequently, interpreting results as reflecting pure compositional effects in the *supply* of labor to the sector hinges on there either being no difference in the way recruits are treated by the education system based on the contract type under which they applied, or on the response being made observable and there being sufficient controls to address any differences in placement. With this in mind, we will consider four demand-side margins on which government might react.

1. *DEOs might select applicants by different criteria in P4P vs FW labor markets.* We will explore this by modeling the probability of (DEO) selection among applicants as a function of applicant characteristics and the interaction of those characteristics with advertised treatment. If we find significant differences across treatments, then by construction it will be on the basis of observed characteristics.[37] Our response will be (a) to consider that our estimates still reflect a policy-relevant quantity—the total effect of advertised P4P on the attributes and performance of placed recruits, net of demand-side responses, and (b) to complement this interpretation with a second set of results that weights placed recruits based on the probability that they would have been hired under a FW advertisement.

[37] Given DEO's limited interactions with teachers, we believe this 'selection on observables' story is a plausible one, and are reassured by the fact that we observe the same CVs as DEOs do.

2. *Differential placement into schools of P4P vs FW recruits.* Even if DEOs select applicants by the same criteria, they may place applicants hired under P4P and applicants hired under FW contracts into systematically different types of schools. We will test for differences in average student performance (at baseline) and in school size, as measured by both the number of students and incumbent teachers in the school at baseline. Under additional assumptions regarding functional form and measurement error, the pre-specified regression controls (e.g., controls for baseline student achievement in the teacher value added regressions) in our primary specification (equation 7) address this potential source of bias. If we find differences in placed school type, we will also estimate weighted regressions that weight advertised P4P outcomes by the probability of placement in a school of that type in the advertised FW arm.

3. *Differential placement into upper-primary teaching.* Once recruits are placed into a school, the head teacher has authority over their teaching responsibilities. Head teacher decisions may differ systematically by advertisement type, either because of the direct effect of the treatment or because they observe a signal of the recruit's ability and condition their classroom assignments on this signal. If we observe significant differences in the placement of teachers into upper-primary classrooms—and so into our main sample—we will complement our primary analysis with estimates that weight placed recruits based on the probability of assignment to upper primary in fixed wage schools.

4. *Differential placement into high- vs low-performing classrooms within school.* This form of non-random placement is a central concern in estimates of value added (e.g., Rothstein, 2010). We will test whether the average ability level of students in the classroom of P4P recruits differs systematically from the average ability level of students in the classroom of FW recruits *conditional on school fixed effects*, both in general and as a function of the realized treatment arm of the school. If we observe significant differences, we will test whether results are consistent when we either (a) restrict attention to any *experienced* treatment arm in which this is not the case, and/or (b) restrict attention to second-year impacts, but use two years of lagged student outcomes as controls.

5.3 Externalities across labor markets

The experiment randomizes contract offers at the district-by-subject level. While the boundaries across subjects of qualification are impervious for potential applicants (who must already be in possession of a TTC degree), district boundaries need not be. We would emphasize that this does not by itself invalidate inference about the existence of a compositional effect, should one be found. But it does matter for the power of such a test, as well as for the policy-relevant magnitude of any compositional effect.

To test for the presence of externalities, we will estimate a model that uses the random assignment of the same teacher qualification in *geographically neighboring* districts to test the hypothesis that the assignment of districts likely to be viewed as close substitutes has an impact on the average quality of the applicant pool in a given district. To do so, we will augment the specification of equation (2) to include geographic neighbors' average treatment status, and we will conduct inference in the manner suggested by Athey et al. (2017).

5.4 Pre-job characteristics of potential applicants

We can use data gathered at the job fairs that we ran at teaching training colleges in December 2015 to help address the limitations that (a) we do not observe the set of *potential* applicants, and

(b) we observe only a limited range of characteristics for applicants who are not placed in schools.

The sample of attendees at our three job fairs is both small and selected on interest, so does not entirely resolve these issues. Nonetheless, we can use the data that we collected at these fairs to model the application decision. Specifically, we will estimate a model of the decision of individual i with qualification q to apply for a job in district d as a function of the advertised treatment status of market qd, a vector of characteristics x_i, and their interaction. Potential applicant characteristics available for this purpose were summarized in Table 6 and include: gender, age, Big Five index, (framed) Dictator Game contribution, choice over a set of risky lotteries, Grading Task score, and decision in the Competition Game. We propose to use a logistic choice model and to test (via randomization inference) the null hypotheses of (a) no effect of Advertised P4P, and (b) no interaction between advertised P4P and the above-mentioned characteristics, on the probability of application.

5.5 Further robustness checks

Additional checks will be undertaken as required to explore the robustness of our primary results. These include:

- *Analysis of advertised and experienced P4P impacts on teacher value added (Hypotheses IV and V) using the panel of pupils tested in multiple rounds*;
- *Analysis of advertised and experienced P4P impacts on teacher value added (HYpotheses IV and V) using Year 1 student outcomes only.* This addresses any possible concerns about the controls for Year 1 outcomes on the right-hand-side of the two-year (pooled) specification. However, both compositional and effort-margin treatment effects may differ across years—for example, as recruits become more experienced their true abilities may emerge—and so we treat any difference between these one-year estimates and our pooled estimate with caution.
- *Analysis of advertised and experienced P4P impacts on teacher value added (Hypotheses IV and V) using Year 2 student outcomes only, and controlling for two lags of stream-mean exam scores*, in the spirit of Rothstein's critique of value added models in the presence of endogenous selection (Rothstein, 2010). This will be compared to Year 2 estimates using only one lag.
- *Testing for differential attrition of recruits between Years 1 and 2.* If evidence is found of differential attrition based on advertised P4P treatment, we will present estimates that address attrition for the primary test of Hypothesis IV; if evidence is found of differential attrition based on experienced P4P, we will present estimates that address attrition for the primary test of Hypothesis V. In both cases, our we will address attrition via inverse-probability weighting.

  We propose to use the (small) panel of pupils in order to estimate a model of attrition probabilities that can be applied to the repeated cross section used in our primary analyses. To do so, we first estimate—among panel pupils j, in streams k, grades g, schools s, and rounds r—a model of the form:

$$\begin{aligned}\Pr(present_{jkgsr}) &= \Phi\left(\tau_A \bar{T}^A_{qd} + \tau_E T^E_s + \tau_{AE}\bar{T}^A_{qd}T^E_s + \lambda_I \bar{I}_j + \lambda_E T^E_s \bar{I}_j + \gamma_g + \delta_d + \psi_r \right. \\ &\quad \left. + f(\text{rank}_{kgsr}(\bar{z}_{j,r-1}))\right). \qquad (10)\end{aligned}$$

  In the notation above, bars above variables denote student-specific means, since any given student may be taught by both recruits and non-recruits, and may be taught by recruits

who applied under different conditions. The function $f(\text{rank}_{kgsr}(\bar{z}_{j,r-1}))$ denotes that the average score (across subjects) of each student in a given stream-grade-school-round $kgsr$ will be ranked, and we will include as the outer function $f(\cdot)$ a second-order polynomial in this rank. Not written for convenience, but important analytically, the f function will also include interactions between this rank term and the full vector of treatment exposures, $\bar{T}^A, T^E, \bar{T}^A T^E, \bar{I}, T^E \bar{I}$.

To apply inverse probability weights to the repeated cross section, we further assume that student ranks are unaffected by their treatment exposures—which, notably, are constant within a stream, since all students in that stream are taught by the same set of teachers in a given year. Under this additional assumption, we can invert the model of equation equation (10) to impute the expected beginning-of-year rank of a student observed at the end of a given year, based on her endline rank and vector of treatments. These expected beginning-of-year ranks will be combined with the vector of treatments to which a student was exposed in order to derive predicted probabilities of presence for IPW estimates.

References

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff, "Teacher turnover, teacher quality, and student achievement in DCPS," CEPA Working Paper No. 16-03 January 2016.

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

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

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

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

Athey, Susan and Guido Imbens, "The econometrics of randomized experiments," in Esther Duflo and Abhijit Banerjee, eds., *Handbook of Economic Field Experiments*, Vol. I, Elsevier, 2017, pp. 73–140.

_ , Dean Eckles, and Guido W Imbens, "Exact p-values for network interference," *Journal of the American Statistical Association*, 2017.

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

Barrett, Garry F and Stephen G Donald, "Consistent tests for stochastic dominance," *Econometrica*, January 2003, *71* (1), 71–104.

Basinga, Paulin, Paul J Gertler, Agnes Binagwaho, Agnes L B Soucat, and Christel M J Vermeersch, "Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation," *Lancet*, April 2011, *377* (9775), 1421–1428.

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

Besley, Timothy and Maitreesh Ghatak, "Competition and Incentives with Motivated Agents," *American Economic Review*, June 2005, *95* (3), 616–636.

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

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

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

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

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

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

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

Clotfelter, Charles, Elizabeth Glennie, Helen Ladd, and Jacob Vigdor, "Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina," *Journal of Public Economics*, 2008, *92*, 1352–1370.

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

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

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

Davidson, Russell and Jean-Yves Duclos, "Statistical inference for stochastic dominance and for the measurement of poverty and inequality," *Econometrica*, November 2000, *68* (6), 1435–1464.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for nothing? Experimental evidence on an unconditional teacher salary increas in Indonesia," *Quarterly Journal of Economics*, 2018, *133* (2), 993–1039.

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

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

Deserranno, Erika, "FInancial incentives as ssignal: Experimental evidence from the recruitment of village promoters in Uganda," Working paper January 2017.

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

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

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

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

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

Figlio, David N and Lawrence W Kenny, "Individual teacher incentives and student performance," *Journal of Public Economics*, 2007, *91*, 901–914.

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

_ , Steven D Levitt, John List, and Sally Sadoff, "Enhancing the efficiency of teacher incentives through loss aversion," NBER Working Paper 18237 2012.

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

Gerber, Alan S and Donald P Green, *Field experiments: Design, analysis, and interpretation*, New York: W. W. Norton & Company, 2012.

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

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

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

Hanushek, Erica A and Steven G Rivkin, "Teacher Quality," in Eric Hanushek and Finis Welch, eds., *Handbook of the Economics of Education*, Vol. 2, Amsterdam: Elselvier B. V., 2006, pp. 1051–1078.

Heathcote, Andrew, Scott Brown, E.J. Wagenmakers, and Ami Eidels, "Distribution-free tests of stochastic dominance for small samples," *Journal of Mathematical Psychology*, October 2010, *54* (5), 454–463.

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

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

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

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

Kahneman, Daniel, Jack L Knetsch, and Richard Thaler, "Fairness as a constraint on profit seeking: Entitlements in the market," *American Economic Review*, September 1986, *76* (4), 728–741.

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

Kane, Thomas J., Daniel F McCaffrey, Trey Miller, and Douglas O Staiger, "Have we identified effective teachers? Validating measures of effective teaching using random assignment," Report of the Measures of Effective Teaching Project, Gates Foundation 2013.

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

Kremer, Michael and Alaka Holla, "Improving education in the developing world: What have we learned from randomized evaluations?," *Annual Review of Economics*, 2009, *1*, 513–542.

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

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

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

Muralidharan, Karthik, "Long-term effects of teacher performance pay: Experimental evidence from India," Unpublished, UCSD April 2012.

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

Neal, Derek, "The Design of Performance Pay in Education," NBER Working Paper No. 16710 January 2011.

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

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

Perry, James L, "Measuring public service motivation: An assessment of construct reliability and validity," *Journal of Public Administration Research and Theory*, January 1996, *6* (1), 5–22.

Perry, James L. and Lois Recascino Wise, "The motivational bases of public service," *Public Administration Review*, 1990, *50* (3), 5–22.

Rockoff, Jonah E, Brian A Jacob, Thomas J Kane, and Douglas O Staiger, "Can you recognize an effective teacher when you hire one?," *Education Finance and Policy*, 2011, *6* (1), 43–74.

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

Rothstein, Jesse, "Student sorting and bias in value added estimation: Selectio on observable and unobservables," *Education Finance and Policy*, Fall 2009, *4* (4), 537–571.

_ , "Teacher quality in educational production: Tracking, decay, and student achievement," *Quarterly Journal of Economics*, February 2010, *125* (1), 175–214.

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

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

Springer, Matthew G, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J R Lockwood, Daniel F McCaffrey, Matthew Pepper, and Brian M Stecher, "Teacher pay for performance: Experimental evidence from the project on incentives in teaching," National Center on Performance Incentives 2010.

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

Woessmann, Ludger, "Cross-country evidence on teacher performance pay," *Economics of Education Review*, 2011, *30*, 404–418.

Young, Alwyn, "Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," Working paper, London School of Economics August 2017.

Appendix A Theory

Section 2 described the details of our two-tiered RCT. In this appendix section, we set out a simple theoretical framework that closely mirrors the RCT. In doing so, we make concrete terms frequently used in the related literature, such as 'teacher skill', 'teacher intrinsic motivation', 'selection', and 'incentives'. Crucially, this framework shows how we can test for compositional (selection) effects—as distinct from incentive effects—using on-the-job teacher performance data.[38]

Appendix A.1 Model

Preferences Individuals are risk neutral and care about compensation w and effort e. Effort costs are sector-specific. Payoffs in the education sector are given by

$$V^E(w, e) = w - e^2 + \tau \cdot e,$$

while payoffs in any other sector are given by

$$V^O(w, e) = w - e^2.$$

Here, we think of $\tau \geq 0$ as *intrinsic motivation*, one dimension of an individual's 'type'. Irrespective of where an individual works, her effort generates a performance metric

$$m = e \cdot \theta + \varepsilon.$$

The parameter $\theta \geq 1$ can be thought of as *ability*, a second dimension of an individual's type.

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

		Advertised	
		FW	P4P
Experienced	FW	a	b
	P4P	c	d

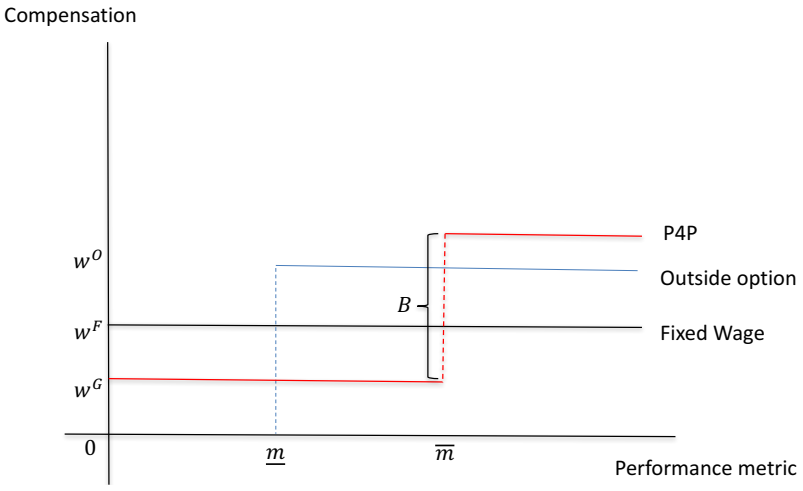
Different compensation schemes are available depending on advertised treamtent status. In the *advertised P4P treatment*, individuals choose between: (i) an education contract of the form, w^G+B if $m \geq \bar{m}$, or w^G otherwise; and (ii) an outside option of the form w^0 if $m \geq \underline{\text{m}}$, or 0 otherwise. In the *advertised FW treatment*, individuals choose between: (i) an education contract of the form w^F; and (ii) the same outside option. In our experiment, the bonus, B, was valued at RWF 100,000, and the fixed-wage contract w^F exceeded guaranteed income in the P4P contract w^G by RWF 20,000. We assume that $w^O > B$ and $w^G + B > w^O > w^F$. The relationship between the performance metric and compensation in the three contractual options is illustrated in Figure A.1.

Timing The timing of the game is as follows.[39]

[38] We stress that we view our theoretical framework largely as a pedagogical device rather than as a means to deliver sharp predictions, and hence one-tailed tests.

[39] For simplicity, we begin by assuming that there is no systematic demand-side selection: employers hire at random. This assumption is not necessary for our predictions about application decisions, provided that application costs are relatively low. However, it may be important to revisit employers' ability to select on type (τ, θ) when thinking about effort responses, which are observed only for hired employees.

Figure A.1: Compensation as a function of the performance metric under alternative contracts



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

Appendix A.2 Analysis

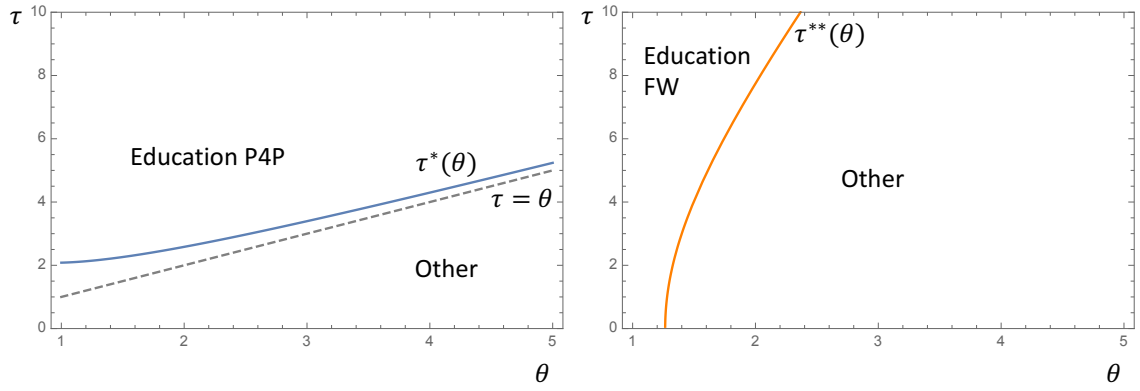
As usual, we solve backwards, starting with effort choices.

Effort incentives Effort choices under the three compensation schemes are:

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

We make two observations. First, intuitively, effort incentives are higher under P4P than under FW. Second, effort in the teacher performance contract, e^P, is only higher than effort in the outside option, e^O, if intrinsic motivation τ is sufficiently high. Notice this result arises because the outside option—perhaps usefully thought of as a private-sector job—has greater wage flexibility than the standard teaching contract: they do not have to pay wages when signals of effort are low. The ordering $B < w^O$ captures the greater stakes in the private-sector contract.

Figure A.2: Decision rules under alternative contract offer treatmetns



Supply-side selection Starting with the advertised P4P treatment, for a given θ, we can define a motivational type τ^* who is indifferent between sectors:

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

Similarly, in the advertised FW treatment, for a given θ, we can define a motivational type τ^{**} who is indifferent between sectors:

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

Figure A.2 illustrates these selection patterns, for a numerical example with $\underline{\varepsilon} = -5$, $\bar{\varepsilon} = 5$, $\underline{m} = 1$, $\bar{m} = 4.5$, $w^O = 50$, $B = 40$, $w^G = 15$. Here, we see a case of *positive selection on intrinsic motivation* and *negative selection on ability* under both treatments. But *less* negative selection on ability under P4P than under FW. Given the single crossing of $\tau^{**}(\theta)$ and $\tau^*(\theta)$ (and distributional assumptions), we have:

$$\tau^P \equiv \mathrm{E}[T|T > \tau^*(\Theta)] > \tau^F \equiv \mathrm{E}[T|T > \tau^{**}(\Theta)]$$
$$\theta^P \equiv \mathrm{E}[\Theta|T > \tau^*(\theta)] > \theta^F \equiv E[\Theta|T > \tau^{**}(\Theta)].$$

In this case, both expected intrinsic motivation and expected ability are higher among P4P applicants than among FW applicants.

Appendix A.3 Empirical implications

Our basic framework formalizes two claims circulating in the survey literature (e.g., Dal Bó and Finan 2016). The first is that P4P creates *incentives*: for a given (τ, θ) type, on-the-job effort is higher under P4P than FW, $e^P > e^F$. The second, and of more relevance for this paper, is that P4P induces *selection* at the recruitment stage: at the time of application, average instrinsic motivation and average skill are higher among individuals recruited under P4P than FW, $\tau^P > \tau^F$ and $\theta^P > \theta^F$.[40]

A central contribution of this paper is that, by virtue of our two-tiered RCT, we can isolate selection in one of our observable measures, namely the incentivised performance metric m. Specifically, using the 2x2 treatment matrix, we can define two selection effects. First, the compositional

[40]It is worth noting that most of the prior literature does not distinguish explicitly between selection at the recruitment stage and selection at the retention stage, as we plan to do across our two papers.

effect of advertised P4P for experienced FW (subgroups a & b):

$$\begin{aligned} E\left[m|\text{advertised P4P, experienced FW}\right] - E\left[m|\text{advertised FW, experienced FW}\right] \\ = e^F(\tau^P) \cdot \theta^P - e^F(\tau^F) \cdot \theta^F. \end{aligned} \tag{11}$$

And second, the compositional effect of advertised P4P for experienced P4P (subgroups c & d):

$$\begin{aligned} E\left[m|\text{advertised P4P, experienced P4P}\right] - E\left[m|\text{advertised FW, experienced P4P}\right] \\ = e^P(\tau^P, \theta^P) \cdot \theta^P - e^P(\tau^F, \theta^F) \cdot \theta^F. \end{aligned} \tag{12}$$

We view the primary role of the theory as pedagogical—to enable us to define these two compositional effects and to show how (power allowing) they can be estimated via a simple comparison of means. The theory also delivers predictions in terms of sign and magnitude of these effects, and they are worth stating here. Specifically, both compositional effects are positive, and the second is larger than the first. If we find that the empirical analogue of (11) and/or (12) is positive, we will conclude that performance contracts *can* attract better teachers, where here a "better" teacher means an individual capable of delivering higher on-the-job teaching performance by virtue of his/her *prior* characteristics.

Although incentive effects are not the focus of this paper, it is helpful to state them here for clarity. Again, there are two effects, an incentive effect of experienced P4P for advertised FW (subgroups a & c):

$$\begin{aligned} E\left[m|\text{experienced P4P, advertised FW}\right] - E\left[m|\text{experienced FW, advertised FW}\right] \\ = e^P(\tau^F, \theta^F) \cdot \theta^F - e^F(\tau^F) \cdot \theta^F. \end{aligned} \tag{13}$$

And an incentive effect of experienced P4P for advertised P4P (subgroups b & d):

$$\begin{aligned} E\left[m|\text{experienced P4P, advertised P4P}\right] - E\left[m|\text{experienced FW, advertised P4P}\right] \\ = e^P(\tau^P, \theta^P) \cdot \theta^P - e^F(\tau^P) \cdot \theta^P. \end{aligned} \tag{14}$$

Appendix B Constructing the 4P performance metric

Our pay-for-performance intervention offers teachers bonuses based on their ranking in their district on a '4P' measure of performance. The performance bonus was awarded to *all* upper primary teachers in schools allocated to the experienced P4P treatment; that is, both the new recruits placed into these schools as well as incumbent teachers already working in these schools.

This performance metric comprises *performance* on student assessments, *presence* of teachers in the school, *preparation* of teachers, and *pedagogy* in the classroom. On each of these components, teachers are ranked within their districts. Their overall score is a weighted average of these component-wise ranks, with the learning outcomes measure receiving half of the weight, and ranks on each of the three input measures (presence, preparation, and pedgagogy) receiving one sixth of the total weight. The subsections below describe the construction of each in turn. First, we briefly explain the rationale behind this 4P metric.

This composite metric was adopted for several reasons. First, it addresses MINEDUC's concern to include direct measures of professional conduct, and to mitigate the risk (via sampling and measurement error) to which a teacher is subjected. Second, it reflects an emergent practice, as in the Gates-funded *Measuring Effective Teaching* study, that such composite metrics can be effective at predicting teachers' contribution to learning (Kane et al., 2013). And third, relative to a pure assessment-based system in which there is little contact between program implementers and teachers over the course of the year, the inclusion of teacher input measures provided a venue for contact that may have increased the salience of the intervention.

The composite metric plays a dual role in our study. First, we used this metric to award bonuses in the experienced P4P arm of the study. Then, since we want to test whether contractual terms (either advertised or experienced) impact teacher behavior, we constructed the same metric in the experienced FW arm of the study, re-calculating ranks using the full sample of available teachers/schools.

Appendix B.1 Performance on student assessments

At the core of our teacher evaluation metric is a measure of the learning gains that teachers bring about, measured by their students' performance on assessments. (See Section 3 for a description of assessment procedures; throughout, we use students' IRT-based predicted abilities to capture their learning outcomes in a given subject and round.) To address concerns over dysfunctional strategic behavior, our objective was to follow Barlevy and Neal's *pay-for-percentile* scheme as closely as was practically possible (Barlevy and Neal, 2012, henceforth BN).

The logic behind the BN scheme is that it creates a series of 'seeded tournaments' that incentivize teachers to promote learning gains at all points in the student performance distribution. In short, a teacher expects to be rewarded equally for enabling a weak student to outperform his/her comparable peers as for enabling a strong student to outperform his/her comparable peers. Roughly speaking, the implemented BN scheme works as follows. Test all students in the district in each subject at the start of the year. Take student j in stream k for subject b at grade g and find that student's percentile rank in the district-wide distribution of performance in that subject and grade at baseline. Call that percentile (or interval of percentiles if data is sparse) student j's baseline bin.[41] Re-test all students in each subject at the end of the year. Establish student

[41] In setting such as ours where the number of students is modest, there is a tradeoff in determining how wide to make the percentile bins. As these become very narrowly defined, they contain few students, and the potential for measurement error to add noise to the results increases. But larger bins make it harder for teachers to demonstrate learning gains in cases where their students start at the bottom of a bin. In practice, we use vigintiles of the

j's end-of-year percentile rank within the comparison set defined by his/her baseline bin. This metric constitutes student j's contribution to the performance score of the teacher who taught that subject-stream-grade that school year. Repeat for all students in all subjects/streams/grades taught by that teacher in that school year, and take the average to give the BN performance metric at teacher level.

We adapt the student test score component of the BN scheme to allow for the fact that we observe only a sample of students in each round in each school-subject-stream-grade. (This was done for budgetary reasons and is a plausible feature of the cost-effective implementation of such a scheme at scale, in an environment in which centrally administered standardized tests are not otherwise taken by all students in all subjects.) To avoid gaming behavior—and in particular, the risk that teachers would distort effort toward those students sampled at baseline—we re-sampled (most) students across rounds, and informed teachers in advance that we would do so.[42]

Specifically, we construct *pseudo-baseline bins* as follows. Students sampled for testing at the end of the year are allocated to district-wide comparison bins using empirical CDFs of start of year performance (of different students). To illustrate, suppose there are 20 baseline bins within a district, and that the best baseline student in a given school-stream-subject-grade is in the (top) bin 20. Then the best endline student in the same school-stream-subject-grade will be assigned to bin 20, and will be compared against all other endline students within the district who have also been placed in bin 20.

To guard against the possibility that schools might selectively withhold particular students selected from the exam, all test takers were drawn from beginning-of-year administrative registers of students in each round. Any student who did not take the test was assigned the minimum theoretically possible score. This feature of our design parallels similar incentives to mitigate incentives for selective test-taking in Glewwe et al. (2010).

Denoting by $z_{j,b,k,g,d,r}$ the IRT estimate of the ability of student j in subject b, stream k, grade g, district d, and round r, with the lowest possible learning level assigned to students who were sampled to take the test but did not do so, we can outline the resulting algorithm for producing the student learning component of the assessment score for rounds $r \in \{1, 2\}$ in the following steps:

1. *Create baseline bins.*

    - Separately for each subject and grade, form a within-district ranking of the students sampled at round $r-1$ on the basis of $z_{j,b,k,g,d,r-1}$. Use this ranking to place these round $r-1$ students into B baseline bins.
    - For each subject-grade-school-stream within a school, calculate the empirical CDF of these baseline bins.[43]

2. *Place end-of-year students into pseudo-baseline bins.*

    - Form a within subject-stream-grade-school percentile ranking of the students sampled at round r on the basis of $z_{j,b,k,g,d,r}$. (In practice, numbers of sampled students varies for a given stream between baseline and endline, so we use percentile ranks rather than simple counts.)

district-subject distribution.

[42] A small panel of students was retained between baseline and the Year 1 endline. For study evaluation purposes, we can in principle revert to using the student panel and hence use genuine baseline bins. We will consider this as a robustness check.

[43] There are 40 subject-grade-school streams (out of a total of 4,175) for which no baseline students were sampled. In such cases, we use the average of the CDFs for the same subject in other streams of the same school and grade (if available) or in the school as a whole to impute baseline learning distributions for performance award purposes.

- Map percentile-ranked students at endline onto baseline bins through the empirical CDF of baseline bins. For example, if there are 20 bins and the best round 1 student in that subject-stream-grade-school was in the top bin, then the best round 2 student in that subject-stream-grade-school, with a percentile rank of 1 for $z_{j,b.k,g,d,2}$, will be placed in pseudo-bin 20.

3. *BN performance metric at student-subject level.* Separately for each subject, grade, and district, form a within-psuedo-baseline bin ranking of the students sampled at round r on the basis of $z_{j,b,g,d,r}$. This is the BN performance metric at student-subject level, which we denote by $\pi_{j,b,k,g,d,r}$. It constitutes student j's contribution to the performance score of the teacher who taught subject b stream k at grade g for school year r.

4. *BN performance metric at teacher-level.* For each teacher, compute the weighted average of the $\pi_{j,b,k,g,t}$ for all the students in the subjects/streams/grades that they taught in round r school year. This is the BN performance metric at teacher-level. Weights w_{jk} are given by the (inverse of the) probability that student j was sampled in stream k: the number of sampled students in that stream divided by the number of students enrolled in the same stream. Note these weights are determined by the number of students *sampled* for the test, *not* the number of students who actually took the test (which may be smaller).[44]

To construct the BN performance metric at teacher-level for the 2017 performance round, $r = 2017$, we must deal with a further wrinkle, namely the fact that we did not sample students at the start of the year. We follow the same procedure as above except that at Step 2 we use the set of students who were sampled for and actually sat the 2016 endline exam, and can be linked to an enrollment status in a specific stream in 2017, to create the baseline bins and CDFs for that year.

Appendix B.2 Input measures

Armed with the BN performance metric at teacher-level, we combine with the input criteria to construct our overall composite metric as follows. Measures of presence (observed through spot checks, with two spot checks in Year 1 and one spot check in Year 2), preparation (the presence of a lesson plan), and pedagogy (summary score based on the Danielson (Danielson, 2007) Framework for Teaching rubric) were defined and described in Section 3. On each of these components, teachers were ranked within their district. The average of the ranks on these three components was weighted equally with the BN learning metric to comprise the teacher's overall performance score.

[44] Our endline sampling frame covered all grades, streams, and subjects. In practice, out of 4,200 school-grade-stream-subjects in the P4P schools, we have data for a sample of students in all but five of these, which were missed in the examination.

Appendix C Power by simulation

This section describes how simulations were derived for estimates of the statistical power of the study for key primary outcomes. We focus in particular on simulations that motivate choice of two specifications, where choices made are not simple (regression-based) tests in means, and where the choices made are motivated by simulation exercises that made use of the blinded data.

These choices are:

- Choice of the specification in Equation (1) as the primary test of Hypothesis I; and
- Choice of the specification in Equation (6) as the basis for primary tests of Hyotheses IV and V.

We discuss them in turn below.

Appendix C.1 Power for analysis of impacts on application quality

In this section, we consider randomization inference-based tests for treatment effects using the following test statistics:

1. the *Kolmogorov-Smirnov statistic*, a test for equality in distributions, which is defined as where $\hat{F}_W(y)$ is the empirical cumulative distribution function of outcome y in treatment arm W, evaluated at some specific value y. As argued in Imbens and Rubin, "because $[T^{KS}]$ is a scalar function of the vector of assignments and the vector of observed outcomes, it is a valid test statistic [for use in randomization inference]." ((Imbens and Rubin, 2015, p. 70)).

2. a *one-sided Kolmogorov-Smirnov statistic*, defined as

$$T^{OKS} = \sup_y \left(\hat{F}_{P4P}(y) - \hat{F}_{FW}(y) \right) = \max_{i=1,\ldots,N} \left(\hat{F}_{FW}(y_i) - \hat{F}_{P4P}(y_i) \right). \tag{15}$$

   which provides a test of the specific hypothesis that F_{P4P} first-order stochastically dominates F_{FW}.[45] Notice this differs from the test statistic T^{KS} in that, for any realized value y_i in the data, it does not take the absolute value of the difference in cumulative distributions at that point. This gives us greater power for the specific alternative hypothesis implied by theory.

3. test for difference in means, implemented as a (two-sided) t statistic.

4. a test for differences in medians, implemented as a (one-sided) Wilcoxon signed-rank statistic.

In addition, we consider alternative mechanisms to 'residualize' these distributions to account for differences that are inherent to either districts or to subjects—where the relevant differences may be differences in either mean or in spread.

To understand the power of these test statistics against specific alternative hypotheses, we undertake the following simulation exercise. Let y_0 denote applicants' true TTC scores, which are reported as percentiles. We scale these between zero and one, with larger values implying better scores. In the simulation, these will be treated as their TTC scores under the FW treatment. Define x_0 as the associated log odds ratio, such that

$$y_0 = \frac{\exp(x_0)}{1 + \exp(x_0)} \tag{16}$$

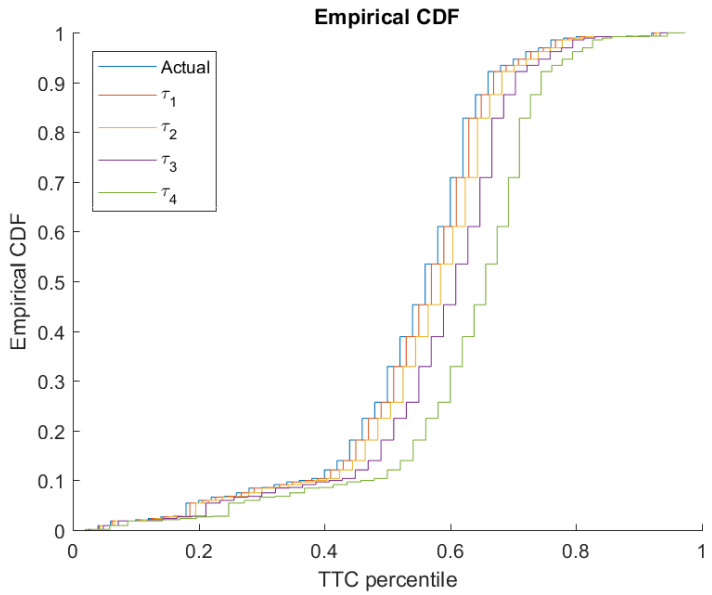
[45] See Davidson and Duclos (2000), Barrett and Donald (2003), and Heathcote et al. (2010) for discussion of power for this and alternative test statistics for first-order stochastic dominance.

Then we define the TTC outcome that would be observed in the presence of treatment as

$$y_1 = \frac{\exp(x_0 + \tau)}{1 + \exp(x_0 + \tau)} \tag{17}$$

for treatment effects $\tau \in \mathcal{T}$. We choose the set of simulated treatments, $\mathcal{T}$, such that they correspond to impacts of 1, 2.5, 5, and 10 percentile points for an applicant whose counterfactual TTC score, y_0, is at the 50th percentile.

Figure C.3: Actual and simulated TTC score distributions for alternative treatment effect sizes



The distributions of these actual and counterfactual TTC score distributions are illustrated in Figure C.3. These potential outcomes are generated for each applicant. For each of $S = 2,000$ simulations, we draw a feasible treatment assignment vector and compute the relevant test statistic for the potential outcomes that would be observed under that assignment, according to the usual switching regression that $y_i = y_{1i}T_i^A + y_{0i}(1 - T_i^A)$, with T_i^A an indicator taking a value of one if the applicant pool was assigned to P4P, and zero if it was assigned to FW. These test statistics are compared with the distribution arising under the sharp null hypothesis that $y_{1i} = y_{0i}$ for all i.

These simulations differ from the planned analysis in two, related ways. First, we have not residualized TTC scores for differences in quality by district or subject (within which the randomization was blocked), as discussion in Section 4.1. We propose to residualize these by linear regression in the final analysis. Second, we make no use of the mixed-treatment applicant pools in these simulations, whereas—as discussed in Section 4.1, we plan to use mixed-treatment applicant pools in the residualization regressions.

We report the power of each test at a significance level of $\alpha = 0.05$ for the values of $\tau \in \mathcal{T}$ described above in Table C.1. As the table reflects, these test statistics appear well powered for effect sizes of the sort simulated here. The one-sided KS statistic is substantially better powered for effect sizes that are very small—a marginal effect of 1 percentage point for an individual at the 50th percentile—but even the two-sided test achieves very strong power when this effect size rises to 2.5 percentage points.

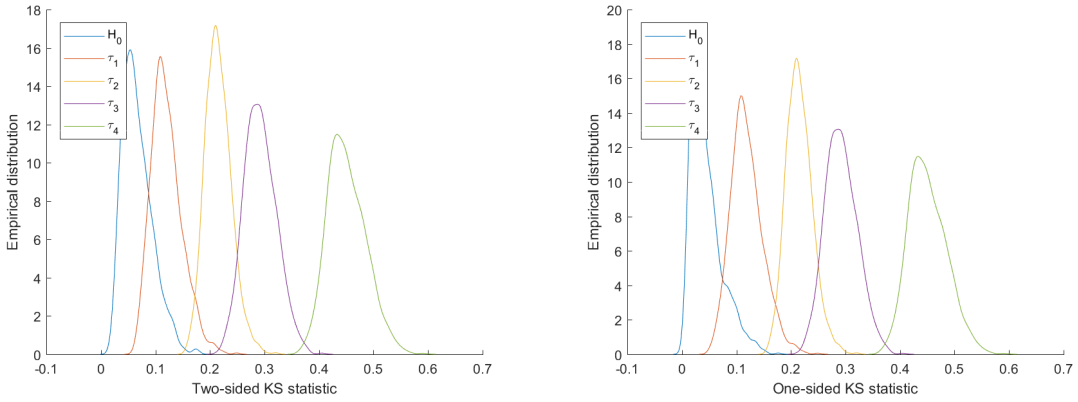
Table C.1: Simulated power of alternative test statistics

Test statistic	τ_1	τ_2	τ_3	τ_4
T^{KS}	0.45	1.00	1.00	1.00
T^{OKS}	0.71	1.00	1.00	1.00
T^{OLS}	0.11	0.37	0.92	1.00
$T^{Wilcoxon}$	0.36	0.98	1.00	1.00
T^{CvM}	0.29	1.00	1.00	1.00

Note: Table reports share of simulated treatments for which sharp null is rejected, for hypothesized values of the treatment effect, τ, such that an individual at the 50th percentile is moved 1, 2, 5, or 10 percentage points, respectively.

The basis for this is seen starkly in the distribution of the relevant test statistics in Figures C.4a and C.4b, which show the distribution of the corresponding test statistics under the sharp null and under each of the simulated treatment values. Notably, power gains of the KS-based (two and one-sided) test statistics relative to inference based on either OLS coefficients or Wilcoxon (signed) rank statistics are dramatic for small effect sizes.

Figure C.4: Simulated power of two- and one-sided KS statistics



(a) Two-sided KS statistic

(b) One-sided KS statistic

Given our desire to be sensitive to a variety of potential changes in the distribution of applicant quality, the marginal power gains from using a test statistic that is sensitive only to violations of equality that imply stochastic dominance seems an overly restrictive tradeoff, not worth making. We conclude that the two-sided KS statistic provides the best omnibus test for changes in distributions in our study. This is reflected in the specification choice of Section 4.1.

We now consider whether, and if so how, to residualize the TTC exam-score distribution prior to the KS test. We consider three alternatives: use the raw TTC scores, use TTC scores residualized using an OLS regression, or use TTC scores residualized within a logit-like transformation.[46] To test consequences for power, we simulated 200 different treamtent assignments. For each of

[46]For the approaches to residualization, we tried doing so using both the full universe of applicants—including mixed-treatment district-subjects—and the subset of district-subjects that were assigned to one of the two 'pure' treatments (for whom the KS test will be employed). Approaches using the sample of pure-treatment markets showed the strongest power, so we include only these in our description here.

Table C.2: Consequences of residualization for power

	τ_1			τ_2			τ_3			τ_4		
Model	raw	OLS	logit	raw	OLS	logit	raw	OLS	logit	raw	OLS	logit
KS	0.342	0.037	0.037	0.998	0.116	0.083	1	0.322	0.037	1	0.584	0.037
One-sided KS	0.45	0.075	0.074	0.999	0.208	0.183	1	0.479	0.074	1	0.691	0.074
OLS	0.099	0.067	0.041	0.294	0.276	0.232	0.758	0.524	0.041	0.999	0.606	0.041
Wilcoxon rank sum	0.227	0.14	0.126	0.901	0.414	0.323	0.996	0.634	0.126	1	0.746	0.126
CVM	0.326	0.069	0.07	0.975	0.25	0.165	1	0.545	0.07	1	0.688	0.07

Note: Each row corresponds to a test statistic. Parameters $\tau_1, \ldots, \tau_4$ represent different simulated treatment effects, using the data generating process and magnitudes in Table C.1. For each such treatment effect, we consider three approaches to residualizing the raw distribution of test scores: none ('raw'), OLS, or logit. The table represents the share of 200 simulations for which we reject the null of equal distributions at the five percent level.

these treamtent assignments, we generated outcomes under treatment effects of size $\tau_1, \ldots, \tau_4$, of magnitudes described in Table C.1 above. Each pair of treatment assignment and effect size constituted a single simulation. Within each of these simulations, we first created raw and residualized TTC exam scores, then calculated the test statistic of interest on the corresponding score. Finally, we calculated these test statistics on each of 1,000 permutations of the treatment effect in order to determine whether the test would reject the sharp null of equal distributions. Reported power figures for each test statistic and effect size are the share of the 200 corresponding trials in which we reject the sharp null.

The OLS residualization included indicators for the full set of districts and subjects. We ran the following regression, for teacher i in district d and qualification q:

$$y_{idq} = \delta_d + \gamma_q + e_{idq} \tag{18}$$

Define $\ddot{e}_{idq}$ as the estimated residual from that regression. Then, defining $\bar{\delta}_d$ as the mean of the district effects and $\bar{\gamma}_q$ as the mean of the qualification effects, we construct $\ddot{y}_{idq} = \bar{\delta}_d + \bar{\gamma}_q + \ddot{e}_{idq}$. This is the outcome measure used in our test for power.

The logit residualization is undertaken as follows. First, we estimate a logistic regression of the form

$$\text{logit}(y_{idq}) = \delta_d + \gamma_q + e_{idq} \tag{19}$$

where $\text{logit}(y_{idq}) = \ln(y_{idq}/(1 - y_{idq}))$. We then combine residuals from this logistic regression with means of disrict and subject effects, and then transform this residualized variable from log-odds space back to the original outcome space, before using it in the testing procedure.

As shown in Table C.2, both approaches to residualization actually *hurt* rather than help statistical power. Consequently, we opt to use the raw TTC scores in our KS test.

Appendix C.2 Power for analysis of teacher value added

The second area where power gains from appropriate specification choice may be substantial is in the estimation of compositional effects of advertised P4P on the value added of recruits placed in schools (Hypothesis IV). As with the application pools, one reason to seek sources of power gain is that the key intervention here occurs at the labor-market (district × qualification) level. Clustering of outcomes within these levels represents a potential threat to power. However, recruits comprise only a small share of the teachers in the schools into which they are placed (typically there are

1–2 recruits in a given school in our sample). Helpful also is the fact we typically think of shocks to learning as occuring at the school, grade, or even student level, but the subject-specialization of teachers means that the outcomes of students in streams and subjects taught by other teachers may be informative about such 'random effects'. To learn about specifications that can utilize this information effectively, we use blinded data that (a) mask the treatment status of recruis and (b) hide entirely one arm of the realized contracts of the study, in order to evaluate the precision of estimates under the 'sharp null' of no treatment effect in Monte Carlo simulations where this is imposed.

Consider the following alternative estimating equations, for the learning outcome of student j in subject b, stream k, school s, district d, and round r:

$$z_{jbksr} = \tau_A T^A_{qd} + \tau_E T^E_s + \tau_{AE} T^A_{qd} T^E_s + \rho_b \bar{z}_{ks,r-1} + \delta_d + e_{jbksr} \quad (20)$$

$$z_{jbksr} = \tau_A T^A_{qd} + \tau_E T^E_s + \tau_{AE} T^A_{qd} T^E_s + \lambda_I I_i + \lambda_E T^E_s I_i + \rho_b \bar{z}_{ks,r-1} + \delta_d + e_{jbksr} \quad (21)$$

As in Section 4, we suppress the dependence of the advertised treatment on the identity of the teacher, i, who is assigned to teach subject-stream pair bk (so that $q = q(i(bk))$). Equation (20) is suitable for estimation using the sample of recruits only. On the other hand, by including an indicator, I_i, for whether the teacher i is an incumbent, Equation (21) permits use of *all* student in the estimating sample.

As written, the advantages of incorporating students of incumbents in a linear regression are minimal; there may be slight differences arising from more precise estimation of district indicators, δ_d, and the ancova coefficients on lagged test scores, ρ_b. However, this setup also permits a richer array of estimators, including fixed-effect estimators (with fixed effects defined at the school level, or below), or random-effects estimators (with random-effects defined likewise). While any estimator that includes fixed effects at the school level or below will not identify the treatment effect on the experienced treatment, τ_E, random-effects estimators have the added advantage of returning an estimate for that parameter as well. In fact, one could argue that this scenario—with school-level errors that are orthogonal to the experimental treatment of interest—is precisely the kind of scenario for which random-effects estimators are designed. But ultimately, we take existence of such power gains will be an empirical question.

To establish the statistical power of alternative specifications, we undertake the following Monte Carlo exercise.

- *Bootstrap sampling of schools.* Under our blinding scheme, we have access to endline data in only one arm of the experienced contract randomization: the 85 schools of the experienced P4P arm. So, for each of B bootstrap iterations, we draw a bootstrap sample of 79 FW schools out of these P4P schools. We repeat this exercise $B = 20$ times to reflect sampling uncertainty induced by the blinding, though the randomization inference estimator that we will use on the full, unblinded sample will reflect uncertainty due to unobserved counterfactuals and random assignment of the treatment, and not due to sampling variation, as discussed in Athey and Imbens (2017).

- *Random (re-)assignment of treatment.* Within each of the B bootstraps, we conduct $P = 200$ random permutations of the treatment vector, within the set of feasible treatment assignments under the original assignment rule.

- *Estimation of model coefficients.* Though our analysis specifies that we will use t statistics as the basis for statistical inference, we find it more helpful to demonstrate the demonstrate the distribution of estimated effect sizes under the Fisher sharp null of no treatment effect. (This

allows comparison to estimated impacts of P4P contracts on the effort margin for incumbent teachers in the literature.) For each of the $B \times P$ samples and randomizations, we estimate the relevant model and collect the key coefficients, τ_A, τ_E, and τ_{AE}, as well as a pooled estimate of the impact of advertised P4P, which we obtain by re-estimating the relevant model without the interaction term, $T^A_{qd} T^E_s$. For clarity, we refer to this pooled effect of advertised P4P as τ^P_A

We consider OLS, random effects, and fixed-effects models, which we report in Table C.3. We also estimate a family of linear mixed-effects models, which impose normality in the distribution of the random effects and error term to estimate coefficients by maximum likelihood.[47] Not shown here, but estimated similarly, are results for an OLS model that omits incumbents altogether, as in equation (20). Since our primary test of Hypothesis IV is based on the poled specification, we focus on the standard deviation of the coefficient τ^P_A in order to choose our preferred specification.

The linear mixed-effects model with (normally distributed) random effects at the round-pupil level performs best by this metric. It has the smallest standard error not only on the pooled estimate of the effect of advertised P4P—our test for a compositional effect on teacher value added—but also performs best in the precision of estimates for the effect of experienced P4P. Indeed, the fact that the coefficient on experienced P4P remains identified in this model represents a considerable advantage over, e.g., the fixed-effects family of models considered.

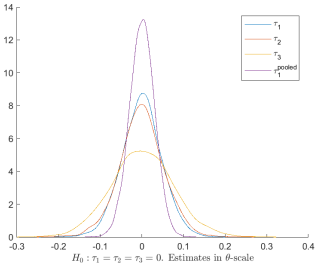
[47] As Imbens and Rubin (2015) note, such a model need not be correctly specified in order to provide a valid test of the sharp null, and the closer it is to 'correct' the greater the power gains that it offers. In this case, the IRT model from which student abilities are derived implies a normal distribution that is consistent with the linear mixed effects model.

Table C.3: Power for alternative test-score specifications: IRT scores

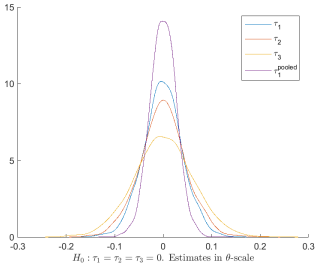
					Distribution under sharp null				
Model	Sample	FE	RE	$\tilde{z}_0$	τ_A	τ_E	τ_{AE}	τ_A^P	$B \cdot P$
OLS models (fixed-effects for dummy variables)									
OLS:D	All	Districts	·	$\bar{z}_{r-1}$	-0.000 (0.048)	-0.000 (0.053)	0.001 (0.075)	0.000 (0.030)	$20 \cdot 200$
Random effects models									
RE:RS	All	Districts	Round-School	$\bar{z}_{r-1}$	-0.001 (0.041)	0.000 (0.048)	0.001 (0.061)	-0.000 (0.027)	$20 \cdot 200$
RE:RK	All	Districts	Round-Stream	$\bar{z}_{r-1}$	0.000 (0.061)	0.000 (0.066)	0.000 (0.092)	0.001 (0.040)	$20 \cdot 200$
RE:RI	All	Districts	Round-Pupil	$\bar{z}_{r-1}$	-0.000 (0.053)	-0.000 (0.058)	0.001 (0.080)	0.001 (0.035)	$20 \cdot 200$
Fixed effects models									
FE:S	All	Schools	·	$\bar{z}_{r-1}$	-0.000 (0.042)	NaN (NaN)	0.000 (0.062)	0.000 (0.029)	$20 \cdot 200$
FE:RS	All	School-Round	·	$\bar{z}_{r-1}$	-0.001 (0.041)	NaN (NaN)	0.001 (0.061)	-0.000 (0.028)	$20 \cdot 200$
FE:RK	All	Stream-Round	·	$\bar{z}_{r-1}$	-0.001 (0.038)	NaN (NaN)	0.001 (0.051)	-0.000 (0.029)	$20 \cdot 200$
FE:RI	All	Student-Round	·	$\bar{z}_{r-1}$	-0.001 (0.045)	NaN (NaN)	-0.000 (0.060)	-0.000 (0.033)	$20 \cdot 200$
Linear mixed-effects models									
LME:S	All	Districts	School	$\bar{z}_{r-1}$	-0.000 (0.042)	0.000 (0.048)	0.001 (0.062)	0.000 (0.029)	$20 \cdot 200$
LME:RS	All	Districts	Round-School	$\bar{z}_{r-1}$	-0.001 (0.041)	0.000 (0.048)	0.001 (0.061)	-0.000 (0.027)	$20 \cdot 200$
LME:RK	All	Districts	Round-Streams	$\bar{z}_{r-1}$	-0.000 (0.039)	0.000 (0.046)	0.000 (0.060)	-0.000 (0.025)	$20 \cdot 200$
LME:RJ	All	Districts	Round-Pupil	$\bar{z}_{r-1}$	-0.000 (0.039)	0.000 (0.044)	0.000 (0.058)	-0.000 (0.025)	$20 \cdot 200$
LME:RS×RK	All	Districts	Round-School, Round-Stream	$\bar{z}_{r-1}$	-0.000 (0.039)	0.000 (0.045)	0.000 (0.056)	-0.000 (0.027)	$20 \cdot 200$
LME:RS×RJ	All	Districts	Round-School, Round-Student	$\bar{z}_{r-1}$	-0.001 (0.039)	0.000 (0.045)	0.000 (0.055)	-0.000 (0.028)	$20 \cdot 200$

Notes: Models of type 'OLS' are OLS estimates with fixed effects implemented as dummy variables. Models of type 'RE' are random effects estimators, with any FE implemented as indicator variables. ANCOVA term $\bar{z}_{r-1}$ refers to vector of once-lagged stream means (spanning all subjects). B denotes the number of bootstrap samples drawn; P denotes the number of permutations taken per bootstrap in randomization inference.

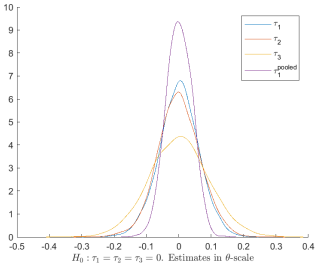
Figure C.5: Distributions of parameter estimates under sharp null: alternative models



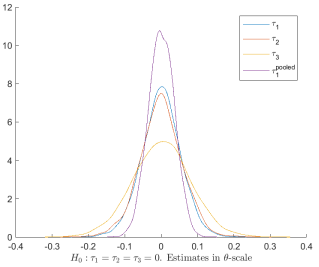
(a) Model OLS:{d}



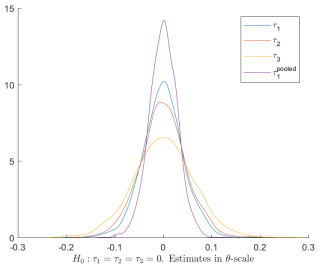
(b) Model RE:{rs}



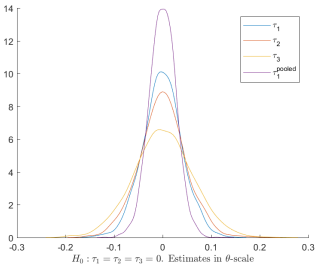
(c) Model RE:{rk}



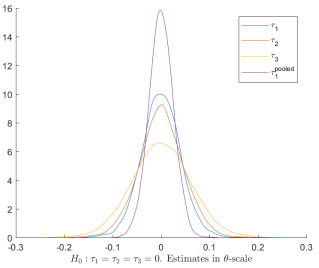
(d) Model RE:{ri}



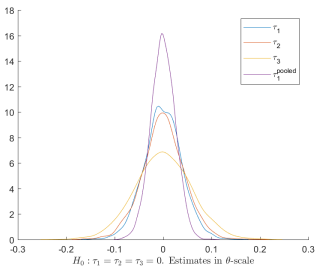
(e) Model LME:{s}



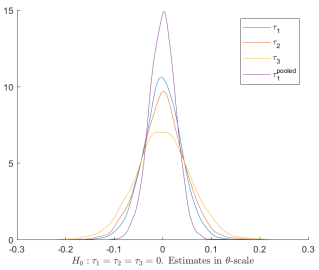
(f) Model LME:{rs}



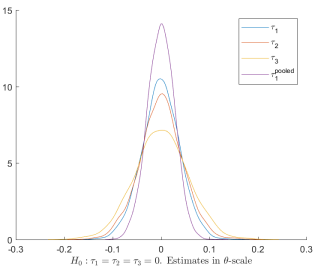
(g) Model LME:{rk}



(h) Model LME:{rj}



(i) Model LME:{rs×rk}



(j) Model LME:{rs×rj}

Finally, because our preferred estimating equation takes the unusual step of controlling for a post-treatment outcome in order to provide an estimate of teacher value added as the object of interest, we undertake a final set of simulations to demonstrate the unbiasedness and power of the resulting estimator when the true (here: imposed) compositional effect of advertised P4P is not zero.

To do so, we estimate a model in which we impose a compositional effect, τ_A, of 0.1 standard

deviations on the value added of recruits who were placed under advertised P4P. We maintain $\tau_E = 0$ for simplicity.

A key question for this analysis is—given that year-one exams are being used as baseline values for year-two outcomes—what share of the year-one treatment effect should persist into year two. On the data-generating process side, we parameterize this persistence, experimenting with values from zero to one, and find that our estimates remain unbiased for τ_A regardless of this value. (On the corresponding analytical side, we allow the ANCOVA coefficient to differ across rounds, in order to allow for the fact that the gap between baseline and endline is longer for year-two outcomes than it is for year-one outcomes.)

Although for computational reasons the number of simulations presented here is smaller, the precision ranking in these results broadly aligns with our prior conclusions. In particular, the linear mixed effects model with random effects at the round-pupil level remains the most precisely estimated for both the effects of advertised and experienced contracts, and the average of 200 coefficient estimates for τ_A^P is within 0.001 of the true (imposed) treatment effect. This confirms our view of the appropriateness of this LME model as the primary basis for our estimates of impacts on teacher value added.

Table C.4: Power and unbiasedness for alternative test-score specifications: IRT scores

					Distribution under alternative hypothesis $\tau_A^P = 0.1$					
Model	Sample	FE	RE	$\tilde{z}_0$	τ_A	τ_E	τ_{AE}	τ_A^P	τ_E^P	$B \cdot P$
OLS models (fixed-effects for dummy variables)										
OLS:D	All	Districts	·	$\bar{z}_{r-1}$	0.097 (0.045)	-0.001 (0.050)	-0.001 (0.071)	0.096 (0.036)	-0.000 (0.026)	$200 \cdot 1$
Random effects models										
RE:RS	All	Districts	Round-School	$\bar{z}_{r-1}$	0.094 (0.054)	-0.000 (0.064)	0.001 (0.087)	0.095 (0.042)	0.001 (0.029)	$200 \cdot 1$
RE:RK	All	Districts	Round-Stream	$\bar{z}_{r-1}$	0.093 (0.055)	-0.000 (0.059)	0.001 (0.083)	0.093 (0.042)	0.001 (0.030)	$200 \cdot 1$
RE:RI	All	Districts	Round-Pupil	$\bar{z}_{r-1}$	0.134 (0.039)	0.036 (0.045)	-0.037 (0.074)	0.095 (0.040)	-0.000 (0.028)	$200 \cdot 1$
Fixed effects models										
FE:S	All	Schools	·	$\bar{z}_{r-1}$	0.101 (0.041)	NaN (NaN)	-0.005 (0.059)	0.098 (0.028)	NaN (NaN)	$200 \cdot 1$
FE:RS	All	School-Round	·	$\bar{z}_{r-1}$	0.100 (0.038)	NaN (NaN)	-0.003 (0.057)	0.099 (0.026)	NaN (NaN)	$200 \cdot 1$
FE:RK	All	Stream-Round	·	$\bar{z}_{r-1}$	0.100 (0.038)	NaN (NaN)	-0.005 (0.052)	0.098 (0.028)	NaN (NaN)	$200 \cdot 1$
FE:RI	All	Student-Round	·	$\bar{z}_{r-1}$	0.103 (0.039)	NaN (NaN)	-0.006 (0.054)	0.100 (0.027)	NaN (NaN)	$200 \cdot 1$
Linear mixed-effects models										
LME:S	All	Districts	School	$\bar{z}_{r-1}$	0.101 (0.041)	0.000 (0.049)	-0.005 (0.059)	0.098 (0.028)	-0.001 (0.033)	$200 \cdot 1$
LME:RS	All	Districts	Round-School	$\bar{z}_{r-1}$	0.100 (0.037)	-0.001 (0.048)	-0.003 (0.057)	0.099 (0.026)	-0.002 (0.034)	$200 \cdot 1$
LME:RK	All	Districts	Round-Streams	$\bar{z}_{r-1}$	0.101 (0.038)	0.002 (0.047)	-0.006 (0.058)	0.098 (0.027)	-0.000 (0.030)	$200 \cdot 1$
LME:RJ	All	Districts	Round-Pupil	$\bar{z}_{r-1}$	0.101 (0.035)	0.001 (0.043)	-0.004 (0.057)	0.099 (0.025)	-0.001 (0.025)	$200 \cdot 1$
LME:RS×RK	All	Districts	Round-School, Round-Stream	$\bar{z}_{r-1}$	0.102 (0.037)	0.000 (0.047)	-0.005 (0.054)	0.099 (0.025)	-0.002 (0.033)	$200 \cdot 1$
LME:RS×RJ	All	Districts	Round-School, Round-Student	$\bar{z}_{r-1}$	0.102 (0.036)	0.000 (0.047)	-0.005 (0.054)	0.099 (0.025)	-0.002 (0.033)	$200 \cdot 1$

Notes: Models of type 'OLS' are OLS estimates with fixed effects implemented as dummy variables. Models of type 'RE' are random effects estimators, with any FE implemented as indicator variables. ANCOVA term $\bar{z}_{r-1}$ refers to vector of once-lagged stream means (spanning all subjects). B denotes the number of bootstrap samples drawn; P denotes the number of permutations taken per bootstrap in randomization inference. Data generated under $\tau_A^p = 0.1$.

Appendix D Supplementary figures and tables

Table D.5: Perry instrument items and subscales

Item	Subscale	Prompt
1	Commitment to the public interest	It is hard for me to get intensely interested in what is going on in my community.
2	Commitment to the public interest	I unselfishly contribute to my community.
3	Commitment to the public interest	Meaningful public service is very important to me.
4	Commitment to the public interest	I consider public service my civic duty.
5	Social justice	I believe that there are many public causes worth championing.
6	Social justice	I am willing to use every ounce of my energy to make the world a more just place.
7	Social justice	I am not afraid to go to bat for the rights of others even if it means I will be ridiculed.
8	Civic duty	I am willing to go great lengths to fulfill my obligations to my country.
9	Civic duty	Public service is one of the highest forms of citizenship.
10	Civic duty	I believe everyone has a moral commitment to civic affairs no matter how busy they are.
11	Civic duty	I have an obligation to look after those less well off.
12	Civic duty	To me, the phrase 'duty, honor, and country' stirs deeply felt emotions.
13	Civic duty	It is my responsibility to help solve problems arising from interdependencies among people.
14	Compassion	I am rarely moved by the plight of the underprivileged.
15	Compassion	It is difficult for me to contain my feelings when I see people in distress.
16	Compassion	To me, patriotism includes seeing to the welfare of others.
17	Compassion	I seldom think about the welfare of people whom I don't know personally.
18	Compassion	I am often reminded by daily events about how dependent we are on one another.
19	Compassion	I have little compassion for people in need who are unwilling to take the first step to help themselves.
20	Self sacrifice	Making a difference in society means more to me than personal achievements.
21	Self sacrifice	I believe in putting duty before self.
22	Self sacrifice	Doing well financially is definitely more important to me than doing good deeds.
23	Self sacrifice	Much of what I do is for a cause bigger than myself.
24	Self sacrifice	Serving citizens would give me a good feeling even if no one paid me for it.
25	Self sacrifice	I feel people should give back to society more than they get from it.
26	Self sacrifice	I am one of those rare people who would risk personal loss to help someone else.
27	Self sacrifice	I am prepared to make enormous sacrifices for the good of society.
28	Compassion	Most social programs are too vital to do without.

Figure D.6: Distribution of responses to components of the Perry PSM instrument among placed recruits in P4P schools