Online Appendix

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

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

Appendix A Supplemental figures and tables

Figure A.1: Study profile

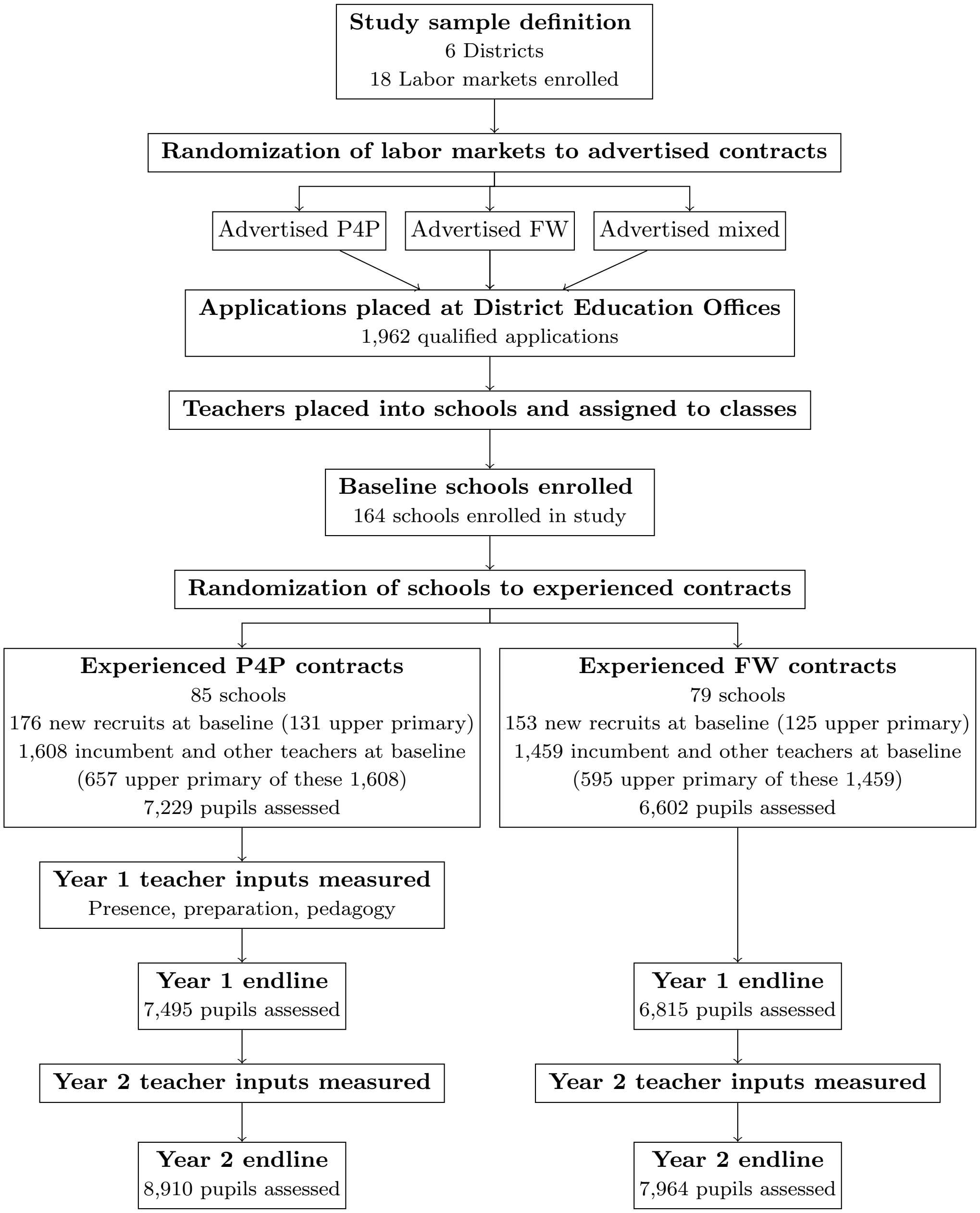
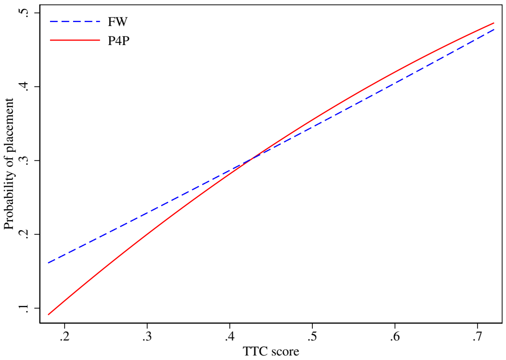

Study sample definition
6 Districts
18 Labor markets enrolled
Randomization of labor markets to advertised contracts
Advertised P4P
Advertised FW
Advertised mixed
Applications placed at District Education Offices
1,962 qualified applications
Teachers placed into schools and assigned to classes
Baseline schools enrolled
164 schools enrolled in study
Randomization of schools to experienced contracts
Experienced P4P contracts
85 schools
176 new recruits at baseline (131 upper primary)
1,608 incumbent and other teachers at baseline
(657 upper primary of these 1,608)
7,229 pupils assessed
Experienced FW contracts
79 schools
153 new recruits at baseline (125 upper primary)
1,459 incumbent and other teachers at baseline
(595 upper primary of these 1,459)
6,602 pupils assessed
Year 1 teacher inputs measured
Presence, preparation, pedagogy
Year 1 endline
7,495 pupils assessed
Year 1 endline
6,815 pupils assessed
Year 2 teacher inputs measured
Year 2 teacher inputs measured
Year 2 endline
8,910 pupils assessed
Year 2 endline
7,964 pupils assessed

Figure A.2: Probability of hiring as a function of TTC score, by advertised treatment arm



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

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

Outcome	Sample	Test statistic	Randomization inference
Hypothesis I: Advertised P4P induces differential application qualities			
*TTC exam scores	Universe of applications	KS test of eq. (**??**)	$\mathcal{T}^A$
District exam scores	Universe of applications	KS test of eq. (**??**)	$\mathcal{T}^A$
TTC exam scores	Universe of applications	t_A in eq. (1)	$\mathcal{T}^A$
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. (1)	$\mathcal{T}^A$
TTC exam scores	Universe of application, weighted by probability of placement	t_A in eq. (1)	$\mathcal{T}^A$
Number of applicants	Universe of applications	t_A in eq. (2)	$\mathcal{T}^A$
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. (**??**)	$\mathcal{T}^A$
Hypothesis III: Advertised P4P induces differentially 'intrinsically' motivated recruits to be placed in schools			
*Dictator-game donations	Placed recruits	t_A in eq. (**??**)	$\mathcal{T}^A$
Perry PSM instrument	Placed recruits retained through Year 2	t_A in eq. (**??**)	$\mathcal{T}^A$
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. (**??**)	$\mathcal{T}^A$
Student assessments	Pooled Year 1 & Year 2 students	t_A and t_{A+AE}; t_{AE} in eq. (**??**)	$\mathcal{T}^A$ $\mathcal{T}^A \times \mathcal{T}^E$

Continues. . .

Table A.1, continued

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

Note: 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 (**??**) and (**??**), which are estimated by maximum likelihood, this is a z rather than t statistic, but we maintain notation to avoid confusion with the test score outcome, z_{jbksr}.

Table A.2: Measures of teacher inputs in P4P schools

	Mean	St Dev	Obs
Year 1, Round 1			
Teacher present	0.97	(0.18)	640
Has lesson plan	0.53	(0.50)	569
Classroom observation: Overall score	2.01	(0.40)	631
Lesson objective	2.00	(0.71)	631
Teaching activities	1.94	(0.47)	631
Use of assessment	1.98	(0.50)	629
Student engagement	2.12	(0.56)	631
Year 1, Round 2			
Teacher present	0.97	(0.18)	629
Has lesson plan	0.53	(0.50)	587
Classroom observation: Overall score	2.27	(0.41)	628
Lesson objective	2.22	(0.76)	627
Teaching activities	2.18	(0.46)	627
Use of assessment	2.23	(0.48)	627
Student engagement	2.46	(0.49)	628
Year 2, Round 1			
Teacher present	0.91	(0.29)	675
Has lesson plan	0.79	(0.41)	568
Classroom observation: Overall score	2.37	(0.34)	520
Lesson objective	2.45	(0.68)	520
Teaching activities	2.28	(0.43)	518
Use of assessment	2.25	(0.47)	519
Student engagement	2.49	(0.45)	520

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

Table A.3: Impacts of advertised P4P on characteristics of placed recruits

	Primary outcomes		Exploratory outcomes			
	Teacher skills	DG contribution	Age	Female	Risk aversion	Big Five
Advertised P4P	-0.184 [-0.836, 0.265] (0.367)	-0.100 [-0.160, -0.022] (0.029)	-0.161 [-1.648, 1.236] (0.782)	0.095 [-0.151, 0.255] (0.325)	0.010 [-0.125, 0.208] (0.859)	-0.007 [-0.270, 0.310] (0.951)
Observations	242	242	242	242	242	241

Note: The table reports the point estimate of τ_A, together with the 95 percent confidence interval in brackets, and the randomization inference p-value in parentheses, from the specification in equation (**??**). The primary outcomes are described in detail in Section **??**. In the third column, the outcome is placed recruit age, measured in years. In the fourth column, the outcome is coded to 1 for female recruits and 0 for males. In the fifth column, the outcome is a binary measure of risk aversion constructed from placed recruits' responses in a hypothetical lottery choice game (Chetan et al., 2010; Eckel and Grossman, 2008). It is coded to 1 when the respondent chooses either of the two riskiest of the five available lotteries, and 0 otherwise (53 percent of the sample make one of these choices). In the final column, the outcome is an index of the Big Five personality traits constructed from the 15 item version, validated by Lang et al. (2011) and following Dohmen and Falk (2010).

Table A.4: Impacts on student learning, OLS model

	Pooled	Year 1	Year 2
Model A: Direct effects only			
Advertised P4P (τ_A)	0.03 [-0.04, 0.14] (0.37)	-0.03 [-0.10, 0.08] (0.51)	0.08 [-0.03, 0.24] (0.10)
Experienced P4P (τ_E)	0.13 [0.03, 0.24] (0.01)	0.10 [0.00, 0.20] (0.05)	0.17 [0.04, 0.32] (0.02)
Experienced P4P × Incumbent (λ_E)	-0.09 [-0.31, 0.15] (0.44)	-0.10 [-0.32, 0.16] (0.40)	-0.09 [-0.34, 0.16] (0.48)
Model B: Interactions between advertised and experienced contracts			
Advertised P4P (τ_A)	0.04 [-0.07, 0.23] (0.41)	-0.03 [-0.14, 0.13] (0.59)	0.12 [-0.03, 0.33] (0.10)
Experienced P4P (τ_E)	0.14 [0.03, 0.26] (0.01)	0.10 [-0.02, 0.22] (0.11)	0.17 [0.02, 0.35] (0.03)
Advertised P4P × Experienced P4P (τ_{AE})	-0.03 [-0.22, 0.17] (0.72)	0.01 [-0.18, 0.21] (0.97)	-0.06 [-0.32, 0.18] (0.60)
Experienced P4P × Incumbent (λ_E)	-0.09 [-0.52, 0.36] (0.62)	-0.09 [-0.47, 0.40] (0.62)	-0.09 [-0.56, 0.51] (0.68)
Observations	154594	70821	83773

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

Table A.5: Teacher endline survey responses

	Job satisfaction	Likelihood of leaving	Positive affect	Negative affect
Model A: Direct effects only				
Advertised P4P	-0.04	-0.07	-0.06	-0.02
	[-0.41, 0.48]	[-0.27, 0.08]	[-0.44, 0.33]	[-0.29, 0.32]
	(0.82)	(0.36)	(0.74)	(0.86)
Experienced P4P	0.05	-0.06	-0.00	0.09
	[-0.25, 0.36]	[-0.18, 0.06]	[-0.28, 0.28]	[-0.14, 0.33]
	(0.72)	(0.39)	(0.99)	(0.47)
Experienced P4P × Incumbent	-0.00	0.04	0.04	-0.07
	[-0.45, 0.48]	[-0.13, 0.21]	[-0.45, 0.52]	[-0.50, 0.37]
	(0.99)	(0.61)	(0.84)	(0.70)
Model B: Interactions between advertised and experienced contracts				
Advertised P4P	-0.10	-0.01	0.02	-0.33
	[-0.57, 0.55]	[-0.26, 0.18]	[-0.52, 0.44]	[-0.75, 0.30]
	(0.67)	(0.93)	(0.89)	(0.20)
Experienced P4P	0.08	-0.07	-0.02	-0.25
	[-0.42, 0.54]	[-0.27, 0.14]	[-0.56, 0.47]	[-0.67, 0.17]
	(0.75)	(0.50)	(0.93)	(0.23)
Advertised P4P × Experienced P4P	0.13	-0.13	-0.16	0.64
	[-0.66, 0.85]	[-0.42, 0.14]	[-0.81, 0.43]	[0.04, 1.28]
	(0.71)	(0.34)	(0.59)	(0.03)
Experienced P4P × Incumbent	-0.03	0.05	0.06	0.27
	[-0.90, 0.90]	[-0.28, 0.37]	[-0.86, 0.90]	[-0.54, 1.09]
	(0.92)	(0.69)	(0.84)	(0.40)
Observations	1483	1492	1474	1447
FW recruit mean (SD)	5.42	0.26	0.31	0.00
	(0.90)	(0.44)	(0.93)	(0.99)
FW incumbent mean (SD)	5.26	0.29	-0.05	0.00
	(1.10)	(0.46)	(1.00)	(1.04)

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

Table A.6: Teacher attitudes toward pay-for-performance at endline

	Very unfavorable	Somewhat unfavorable	Neutral	Somewhat favorable	Very favorable
Recruits applying under FW (64)	4.7%	4.7%	7.8%	10.9%	71.9%
—Experiencing FW (33)	6.1%	9.1%	9.1%	3.0%	72.7%
—Experiencing P4P (31)	3.2%	0.0%	6.5%	19.4%	71.0%
Recruits applying under P4P (60)	5.0%	3.3%	8.3%	1.7%	81.7%
—Experiencing FW (32)	6.2%	0.0%	6.2%	0.0%	87.5%
—Experiencing P4P (28)	3.6%	7.1%	10.7%	3.6%	75.0%
Incumbent teachers (1,113)	5.0%	7.5%	7.2%	9.9%	70.4%
—Experiencing FW (537)	5.2%	8.6%	8.0%	8.6%	69.6%
—Experiencing P4P (576)	4.9%	6.6%	6.4%	11.1%	71.0%

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

Appendix B Theory

This appendix sets out a simple theoretical framework, adapted from Leaver et al. (2019), that closely mirrors the experimental design described in Section **??**. We used this framework as a device to organize our thinking when choosing what hypotheses to test in our pre-analysis plan. We did not view the framework as a means to deliver sharp predictions for one-tailed tests.

The model

We focus on an individual who has just completed teacher training, and who must decide whether to apply for a teaching post in a public school, or a job in a generic 'outside sector'.[1]

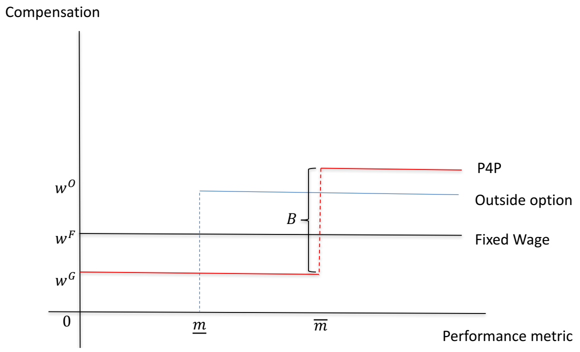
Preferences The individual is risk neutral and cares about compensation w and effort e. Effort costs are sector-specific. The individual's payoff in the education sector is $w-(e^2-\tau\, e)$, while her payoff in the outside sector is $w-e^2$. The parameter $\tau \geq 0$ captures the individual's *intrinsic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

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

Compensation schemes Different compensation schemes are available depending on advertised treatment status. In the advertised P4P treatment, individuals choose between: (i) an education contract of the form, $w^G + B$ if $m \geq \overline{m}$, or w^G otherwise; and (ii) an outside option of the form w^0 if $m \geq \underline{m}$, or 0 otherwise. In the advertised FW treatment, individuals choose between: (i) an education contract of the form w^F; and (ii) the same outside option. In our experiment, the bonus B was valued at RWF 100,000, and the fixed-wage contract exceeded the guaranteed income in the P4P contract by RWF 20,000 (i.e. $w^F - w^G = 20,000$).

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

Figure B.1: Compensation schemes in the numerical example



Timing The timing of the game is as follows.

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

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

Analysis

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

Effort incentives Effort choices under the three compensation schemes are:

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

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

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

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

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

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

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

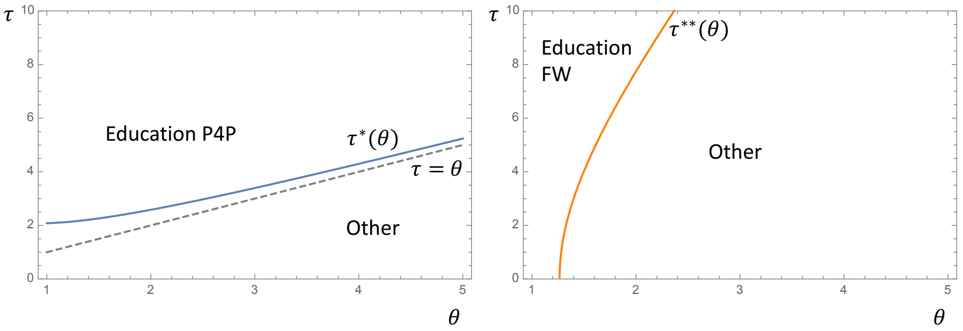
Empirical implications

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

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

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

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



and

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

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

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

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

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

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

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

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

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

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

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

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

Appendix C Applications

Here, we report results from secondary tests of Hypothesis I: advertised P4P induces differential application qualities, and also provide a robustness check of our assumption that district-by-subject-family labour markets are distinct.

Secondary tests

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

$$y_{iqd} = \tau_A T^A_{qd} + \gamma_q + \delta_d + e_{iqd}, \quad \text{with weights } w_{iqd}, \tag{1}$$

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

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

where q indexes subject families and d indexes districts; N_{qd} measures the number of qualified applicants in each district.[2] Although our pre-analysis plan proposes a fifth test—a KS test of equation (**??**) using district exam scores—we did not do this because our sample of these scores was incomplete.

To undertake inference about these differences in means, we use randomization inference, sampling repeatedly from the set of potential (advertised) treatment assignments $\mathcal{T}^{\mathcal{A}}$. Following Chung and Romano (2013), we studentize this parameter by dividing it by its (cluster-robust, clustered at the district-subject level) standard error to control the asymptotic rejection probability against the null hypothesis of equality of means. These are two-sided tests.[3] The absolute value of the resulting test statistic, $|t_A|$, is compared to its randomization distribution in order to provide a test of the hypothesis that $\tau_A = 0$.

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

[2] '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.

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

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

	KS	Unweighted	Empirical weights	Top	Number of Applicants
Advertised P4P	n.a. [-0.020, 0.020] (0.909)	-0.001 [-0.040, 0.036] (0.984)	-0.001 [-0.038, 0.032] (0.948)	-0.009 [-0.025, 0.008] (0.331)	-0.040 [-0.306, 0.292] (0.811)
Observations	1715	1715	1715	1715	18

Note: The first column shows the confidence interval in brackets, and the p-value in parentheses, from the primary KS test discussed in Section **??**. The second column reports the (unweighted OLS) point estimate of τ_A from the applicant TTC exam score specification in (1). The third and fourth columns report the point estimate of τ_A from the same specification with the stated weights. The fifth column reports the point estimate of τ_A from the number of applicants per labor market specification in (2), with the outcome N_{qd} in logs.

randomization inference p-value is 0.984, indicating that we cannot reject the sharp null of no impact of advertised P4P. The third column reports results for the TTC score regression with weights $w_{iqd} = \hat{p}_{iqd}$, where $\hat{p}_{iqd}$ is the estimated probability of being hired as a function of district and subject indicators, as well as a fifth-order polynomial in TTC exam scores, estimated using FW applicant pools only (i.e. the status quo mapping from TTC scores to hiring probabilities). The fourth column reports results for the TTC score regression with weights $w_{iqd} = 1$ for the top $\hat{H}$ teachers in their application pool, and zero otherwise (i.e. a meritocratic hiring rule based on TTC scores alone). Here, we test for impacts on the average ability of the top $\hat{H}$ applicants, where $\hat{H}$ is the predicted number hired in that district and subject based on outcomes in advertised FW district-subjects. Neither set of weights changes the conclusion from the second column: we cannot reject the sharp null of no impact of advertised P4P. The final column reports results for the (logged) application volume regression. Our estimate of τ_A is -0.040. The randomization inference p-value of 0.811, indicating that we cannot reject the sharp null of no impact of advertised P4P on application volumes.

Robustness

To illustrate the implications of cross-district applications, consider an individual living in, say, Ngoma with the TTC qualification of TSS. On the assumption that this individual is willing to travel only to the neighbouring district of Rwamagana, she could be impacted by the contractual offer of P4P in her home 'Ngoma-TSS' market and/or the contractual offer of P4P in the adjacent 'Rwamagana-TSS' market. That is, she might apply in both markets, or in Rwamagana instead of Ngoma—what we term *a cross-district labor-supply effect.* The former behavior would simply make it harder to detect a selection effect at the application stage (although not at the placement stage since only one job can be accepted). But the latter cross-district labor-supply effect would be more worrying. We would not find a selection effect where none existed—without a direct effect of advertised P4P on a given market, there cannot be cross-district effects by this posited mechanism—but we might overstate the magnitude of any selection effect.

Our random assignment provides us with an opportunity to test for the presence cross-district labor-supply effects. To do so, we construct an *adjacency matrix*, defining two labor markets as adjacent if they share a physical border and the same TTC subject-family qualification. We then construct a count of the number of adjacent markets that are assigned to Advertised P4P, and an analogous count for 'mixed' treatment status. Conditional on the number of adjacent markets, this measure of the local saturation of P4P is randomly determined by the experimental assignment of districts to advertised contractual conditions. A regression of labor-market outcomes in a given district on both its own advertised contractual status (direct effect) and this measure of local saturation, conditional on the number of neighboring labor markets, provides an estimate of cross-district labor-supply effects and, by randomization inference, a test for their presence.

Table C.2: Cross-district effects in teacher labor market outcomes

	TTC scores	Number of applicants
Advertised P4P	0.032 [-0.050, 0.103] (0.297)	-0.085 [-0.469, 0.972] (0.900)
Adjacent P4P markets	0.027 [-0.022, 0.087] (0.115)	-0.047 [-0.833, 0.573] (0.710)
Observations	1715	18

Note: The table shows point estimates for the direct and local saturation effects of P4P contracts, with confidence intervals in brackets and randomization inference *p*-values in parentheses. In the first column, the unit of analysis is the application and the outcome is the TTC score of the applicant. In the second column, the unit of analysis is the labor market and the outcome is the number of applications, in logs. All specifications control for the total number of adjacent markets.

Table C.2 shows results of this analysis for two key labor-market outcomes—applicant TTC scores analyzed at the application level, and the number of applications per labor-market analyzed at the labor-market level. The direct effects of advertised P4P on each of these outcomes are presented for comparison and remain qualitatively unchanged relative to the estimates in Table C.1, which did not allow for saturation effects. Estimated saturation effects of neighboring P4P markets are modest in estimated size and statistically insignificant for both outcomes. This suggests that saturation effects were of limited consequence in our setting.

Appendix D Test-score constructs

Barlevy-Neal metric

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 **??** 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 i 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 i's baseline bin.[4] Re-test all students in each subject at the end of the year. Establish student i's end-of-year percentile rank within the comparison set defined by his/her baseline bin. This metric constitutes student i'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.

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

[4] 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 district-subject distribution.

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).

Denote by z_{ibkgdr} the IRT estimate of the ability of student i in subject b, stream k, grade g, district d, and round r. 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 $z_{ibkgd,r-1}$. 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.[5]

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_{ibkgdr}. In practice, numbers of sampled students varies for a given stream between baseline and endline, so we use percentile ranks rather than simple counts. Assign the lowest possible learning level to students who were sampled to take the test but did not do so.
    - 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 will be placed in pseudo-bin 20.

[5] 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.

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_{ibgdr}. This is the BN performance metric at student-subject level, which we denote by π_{ibkgdr}. It constitutes student i'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 π_{ibkgt} for all the students in the subject-stream-grades that they taught in round r school year. This is the BN performance metric at teacher-level. Weights w_{ik} are given by the (inverse of the) probability that student i 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), since our implementation of the BN metric includes, with the penalty described above, students who were sampled for but did not sit the test.[6]

To construct the BN performance metric at teacher-level for the second performance round, $r = 2$, 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 round 1 endline exam, and can be linked to an enrollment status in a specific stream round 2, to create the baseline bins and CDFs for that year.

Teacher value added

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

We adapt the approach taken in prior literature, most notably Kane and Staiger (2008) and Bau and Das (2020). Denoting as in equations (**??**) and (**??**) the learning outcomes of student i in subject b, stream k of grade g, taught by teacher j in school s and round r by z_{ibgjsr}, we express the data-generating process as:

$$z_{ibgjsr} = \rho_{bgr}\bar{z}_{ks,r-1} + \mu_{bgr} + \lambda_s + \theta_j + \eta_{jr} + \varepsilon_{ibgjr}, \tag{3}$$

This adapts a standard TVA framework to use the full pseudo-panel of student learning measures. Our sampling strategy implies that most students are not observed in consecutive assessments, as discussed in Section **??**. We proxy for students' baseline abilities using the vector of means of lagged learning outcomes in all subjects, $\bar{z}_{ks,r-1}$, where the parameter ρ_{bgr} allows these lagged mean outcomes to have

[6] 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.

distinct own- and cross-subject associations with subsequent learning for all subjects, grades, and rounds. In a manner similar to including means instead of fixed effects (Chamberlain, 1982; Mundlak, 1978), these baseline peer means block any association between teacher ability (value added) and the baseline learning status of sampled students.

In equation (3), the parameter θ_j is the time-invariant effect of teacher j: her value added. We allow for fixed effects by subject-grade-rounds, μ_{bgr}, and schools λ_s, estimating these within the model. We then form empirical Bayes estimates of TVA as follows.

1. Estimate the variance of the TVA, teacher-year, and student-level errors, $\theta_j, \eta_{jr}, \varepsilon_{ibgjr}$ respectively, from equation (3). Defining the sum of these errors as $v_{ibgjr} = \theta_j + \eta_{jr} + \varepsilon_{ibr}$: the last variance term can be directly estimated by the variance of student test scores around their teacher-year means: $\hat{\sigma}^2_\varepsilon = \text{Var}(v_{ibgjr} - \bar{v}_{jr})$; the variance of TVA can be estimated from the covariance in teacher mean outcomes across years: $\hat{\sigma}^2_\theta = \text{Cov}(\bar{v}_{jr}, \bar{v}_{j,r-1})$, where this covariance calculation is weighted by the number of students taught by each teacher; and the variance of teacher-year shocks can be estimated as the residual, $\hat{\sigma}^2_\eta = \text{Var}(v_{ibgjr}) - \hat{\sigma}^2_\theta - \hat{\sigma}^2_\varepsilon$.

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

3. Construct the empirical Bayes estimate of each teacher's value added by multiplying this weighted average of classroom residuals, $\bar{v}_j$, by an estimate of its reliability:
$$\widehat{VA}_j = \bar{v}_j \left(\frac{\hat{\sigma}^2_\theta}{\text{Var}(\bar{v}_j)} \right) \tag{4}$$
where $\text{Var}(\bar{v}_j) = \hat{\sigma}^2_\theta + \left(\sum_r h_{jr}\right)^{-1}$, with $h_{jr} = \text{Var}(\bar{v}_{jr}|\theta_j)^{-1} = \left(\hat{\sigma}^2_\eta + \frac{\hat{\sigma}^2_\varepsilon}{n_{jr}}\right)^{-1}$.

Following this procedure, we obtain a distribution of (empirical Bayes estimates of) teacher value added for placed recruits who applied under advertised FW. The Round 2 point estimate from the student learning model in Equation (**??**) would raise a teacher from the 50th to above the 76th percentile in this distribution. Figure **??** plots the distributions of (empirical Bayes estimates of) $\theta_j + \eta_{jr}$ separately for $r = 1, 2$, and for recruits applying under advertised FW and advertised P4P.

It is of interest to know whether the measures of teacher ability and intrinsic motivation that we use in Section **??** are predictive of TVA. This is undertaken in Table D.1, where TVA is the estimate obtained pooling across rounds and treatments.[7] Interestingly, the measure of teacher ability that we observe among recruits at baseline, Grading Task IRT score, *is* positively correlated with TVA (rank correlation of 0.132, with a p-value of 0.039). It is also correlated with TTC final exam

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

score (rank correlation of 0.150, with a p-value of 0.029). However, neither the measure that districts have access to at the time of hiring, TTC final exam score, nor the measure of intrinsic motivation that we observe among recruits at baseline, DG share sent, is correlated with TVA.

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

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

Note: The table provides rank correlations and associated p-values (in parentheses) for relationships between recruits' teacher value added and various measures of skill and motivation: TTC final exam scores, baseline Grading Task IRT scores, and baseline Dictator Game share sent. We obtain the empirical Bayes estimate of TVA from θ_j estimated in the school fixed-effects model in equation (3).

Appendix E Communication about the intervention

Promotion to potential applicants

The subsections below give details of the (translated) promotional materials that were used in November and December 2015.

Leaflets and posters in district offices

A help desk was set up in every District Education Office. Staffers explained the advertised contracts to individuals interested in applying, and distributed the leaflet shown in Figure E.1, and stickers. Permanent posters, like the example shown in Figure E.2 further summarised the programme. Staffers kept records of the number of visitors and most frequent questions, and reported back to head office.[8]

Radio Ads

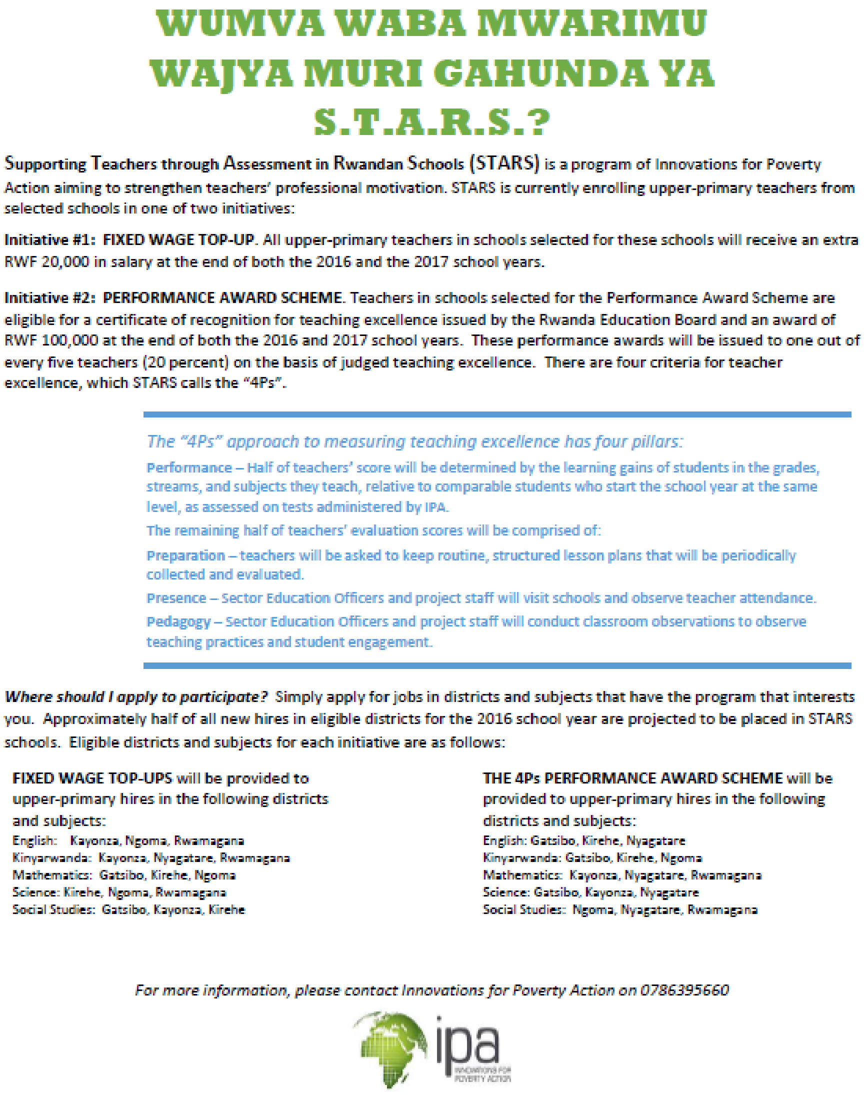
Radio ads were broadcast on Radio Rwanda, the national public broadcaster, during November/December 2015 to promote awareness of the intervention. The scripts below were developed in partnership with a local advertising agency.

Radio script 1 *SFX: Noise of busy environment like a trading centre*

> FVO: Hey, Have you seen how good Gasasira's children look? [*This is a cultural reference implying that teachers are smart, respected individuals and nothing literal about how the child looks.*]
>
> MVO: Yeah! That's not surprising though, their parents are teachers.
>
> FVO: Hahahahah...[*Sarcastic laugh as if to say, what is so great about that.*]
>
> MVO: Don't laugh...haven't you heard about the new programme in the district to recognize and reward good teachers? I wouldn't be surprised if Gasasira was amongst those that have been recognised.
>
> ANNOUNCER: Innovations for Poverty Action in collaboration with REB and MINEDUC, is running the STARS program in the districts Kayonza, Ngoma, Rwamagama, Kirehe, Gatsibo, and Nyagatare for the 2016 academic year. Some *new* teachers applying to these districts will be eligible for STARS which rewards the hardest working, most prepared and best performing teachers. Eligible districts are still being finalized—keep an eye out for further announcements!

[8] The respective number of visitors were: Gatsibo 305, Kayonza 241, Kirehe 411, Ngoma 320, Nyagatara 350, and Rwamagama 447.

Figure E.1: Leaflet advertising treatments



WUMVA WABA MWARIMU WAJYA MURI GAHUNDA YA S.T.A.R.S.?

Supporting Teachers through Assessment in Rwandan Schools (STARS) is a program of Innovations for Poverty Action aiming to strengthen teachers' professional motivation. STARS is currently enrolling upper-primary teachers from selected schools in one of two initiatives:

Initiative #1: FIXED WAGE TOP-UP. All upper-primary teachers in schools selected for these schools will receive an extra RWF 20,000 in salary at the end of both the 2016 and the 2017 school years.

Initiative #2: PERFORMANCE AWARD SCHEME. Teachers in schools selected for the Performance Award Scheme are eligible for a certificate of recognition for teaching excellence issued by the Rwanda Education Board and an award of RWF 100,000 at the end of both the 2016 and 2017 school years. These performance awards will be issued to one out of every five teachers (20 percent) on the basis of judged teaching excellence. There are four criteria for teacher excellence, which STARS calls the "4Ps".

The "4Ps" approach to measuring teaching excellence has four pillars:

Performance – Half of teachers' score will be determined by the learning gains of students in the grades, streams, and subjects they teach, relative to comparable students who start the school year at the same level, as assessed on tests administered by IPA.

The remaining half of teachers' evaluation scores will be comprised of:

Preparation – teachers will be asked to keep routine, structured lesson plans that will be periodically collected and evaluated.

Presence – Sector Education Officers and project staff will visit schools and observe teacher attendance.

Pedagogy – Sector Education Officers and project staff will conduct classroom observations to observe teaching practices and student engagement.

Where should I apply to participate? Simply apply for jobs in districts and subjects that have the program that interests you. Approximately half of all new hires in eligible districts for the 2016 school year are projected to be placed in STARS schools. Eligible districts and subjects for each initiative are as follows:

FIXED WAGE TOP-UPS will be provided to upper-primary hires in the following districts and subjects:

English: Kayonza, Ngoma, Rwamagana
Kinyarwanda: Kayonza, Nyagatare, Rwamagana
Mathematics: Gatsibo, Kirehe, Ngoma
Science: Kirehe, Ngoma, Rwamagana
Social Studies: Gatsibo, Kayonza, Kirehe

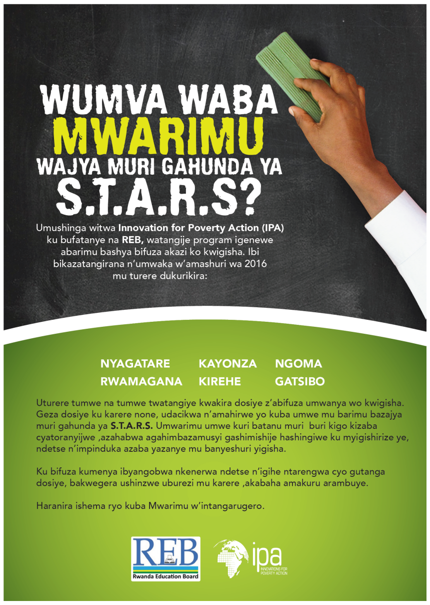
THE 4Ps PERFORMANCE AWARD SCHEME will be provided to upper-primary hires in the following districts and subjects:

English: Gatsibo, Kirehe, Nyagatare
Kinyarwanda: Gatsibo, Kirehe, Ngoma
Mathematics: Kayonza, Nyagatare, Rwamagana
Science: Gatsibo, Kayonza, Nyagatare
Social Studies: Ngoma, Nyagatare, Rwamagana

For more information, please contact Innovations for Poverty Action on 0786395660

ipa
INNOVATIONS FOR POVERTY ACTION

Figure E.2: Poster explaining the programme


WUMVA WABA
MWARIMU
WAJYA MURI GAHUNDA YA
S.T.A.R.S?
REB
Rwanda Education Board
ipa
INNOVATIONS FOR POVERTY ACTION

Umushinga witwa **Innovation for Poverty Action (IPA)** ku bufatanye na **REB,** watangije program igenewe abarimu bashya bifuza akazi ko kwigisha. Ibi bikazatangirana n'umwaka w'amashuri wa 2016 mu turere dukurikira:

NYAGATARE KAYONZA NGOMA
RWAMAGANA KIREHE GATSIBO

Uturere tumwe na tumwe twatangiye kwakira dosiye z'abifuza umwanya wo kwigisha. Geza dosiye ku karere none, udacikwa n'amahirwe yo kuba umwe mu barimu bazajya muri gahunda ya **S.T.A.R.S.** Umwarimu umwe kuri batanu muri buri kigo kizaba cyatoranyijwe ,azahabwa agahimbazamusyi gashimishije hashingiwe ku myigishirize ye, ndetse n'impinduka azaba yazanye mu banyeshuri yigisha.

Ku bifuza kumenya ibyangobwa nkenerwa ndetse n'igihe ntarengwa cyo gutanga dosiye, bakwegera ushinzwe uburezi mu karere ,akabaha amakuru arambuye.

Haranira ishema ryo kuba Mwarimu w'intangarugero.

Radio script 2 *SFX: Sound of a street with traffic and cars hooting*

VO1: Mari, hey Mariko!....What's the rush, is everything OK?

VO2: Oh yes, everything is fine. I am rushing to apply for a job and don't want to find all the places taken.

VO1: Oh that's good. And you studied to be a teacher right?

VO2: Exactly! Now I am going to submit my papers at the District Office and hope I get lucky on this new programme that will be recognizing good teachers!

ANNOUNCER: Innovations for Poverty Action in collaboration with REB and MINEDUC, is running the STARS program in the districts Kayonza, Ngoma, Rwamagama, Kirehe, Gatsibo, and Nyagatare for the 2016 academic year. Some *new* teachers applying to these districts will be eligible for STARS which rewards the hardest working, most prepared and best performing teachers. Eligible districts are still being finalized—keep an eye out for further announcements!

Radio script 3 *SFX: Calm peaceful environment*

VO1: Yes honestly, Kalisa is a very good teacher!

VO2: You are right, ever since he started teaching my son, the boy now understands maths!

VO1: Yes and because of him other parents want to take their children to his school.

VO2: Aaah!...That must be why he was selected for the programme that rewards good teachers.

VO1: He definitely deserves it, he is an excellent teacher.

ANNOUNCER: Innovations for Poverty Action in collaboration with REB and MINEDUC, is running the STARS program in the districts Kayonza, Ngoma, Rwamagama, Kirehe, Gatsibo, and Nyagatare for the 2016 academic year. Some *new* teachers applying to these districts will be eligible for STARS which rewards the hardest working, most prepared and best performing teachers. Eligible districts are still being finalized—keep an eye out for further announcements!

Briefing in P4P schools

The subsections below provide extracts of the (translated) script that was used during briefing sessions with teachers in P4P schools in April 2016. The main purpose of these sessions was to explain the intervention and maximise understanding of the new contract.

Introduction

[Facilitator speaks.] You have been selected to participate in a pilot program that Rwanda Education Board (REB) and Innovations for Poverty Action (IPA) are undertaking together on paid incentives and teacher performance. As a participant in this study, you will be eligible to receive a competitive bonus based on your performance in the study. The top 20 percent of teachers in participating schools in your district will receive this bonus. All participants will be considered for this paid bonus. It is important to note that your employment status will not be affected by your participation in this study. It will not affect whether you keep your job, receive a promotion, etc.

You will be evaluated on **four different categories**:

1. **Presence**, which we will measure through whether you are present in school on days when we visit;

2. **Preparation**, which we will measure through lesson planning;

3. **Pedagogy**, which we will measure through teacher observation; and

4. **Performance**, which we will measure through student learning assessments. You will receive additional information on each of these categories throughout this training.

In your evaluation, the first three categories (presence, preparation, and pedagogy) will contribute equally to your 'inputs' score. This will be averaged with your 'performance' score (based on student learning assessments) which will therefore be worth half of your overall score. [Teachers are then provided with a visual aid.]

The SEO will now tell you how we are going to measure each of these components of your performance. Before I do so, are there any questions?

Presence: Teacher attendance score

[SEO now speaks.] I will now explain to you the first component of your performance score: Teacher Presence. During this pilot program, I will visit your school approximately one time per term. Sometimes I will come twice or more; you will not know in advance how many times I plan to visit in any term. These visits to your school will be unannounced. Neither your Head Teacher nor you will know in advance when I plan to visit your school. I will arrive approximately at the start of the school day. Teachers who are present at that time will be marked 'present'; those that are not will be marked 'late' or 'absent'. The type of absence will be recorded. Teachers who have excused reasons for not being present in school will be marked 'excused'. These reasons include paid leaves of absence, official trainings, and sick leaves that have been granted in advance by the Head Teacher. If you are not present because you feel unwell but have not received advance permission from the Head Teacher, you will be marked as absent.

It is in your best interest to be present every day, or in the case of emergency, notify the head teacher of your absence with an appropriate excuse before the beginning of classes. I will also record what time you arrive to school. You will be marked for arriving on time and arriving late to work. It is in your best interest to arrive on time to school every day.

Preparation: Lesson planning score

Later in this session, you'll be shown how to use a lesson planning form. Lesson planning is a tool to help you improve both your organization and teaching skills. The lesson planning form will help you to include the following components into your lesson:

- A clear lesson objective to guide the lesson.
- Purposeful teaching activities that help students learn the skill.
- Strong assessment opportunities or exercise to assess students' understanding of the skill.

This lesson planning form consists of three categories: lesson objective, teaching activities, assessment/exercises. You will be evaluated on these three categories. I will not evaluate your lesson plans. Instead, I will collect your lesson planning forms at the end of the study. An IPA education specialist will review your lesson plans and score them. They will compare your lesson plans to other teachers' plans in the district. Please be aware that these lesson plans will only be used for this study and will not be reviewed by any MINEDUC officials. They will use the following scoring scale, with 0 being the lowest score and 3 being the highest score. [Teachers are then provided with a visual aid.]

You will be responsible for filling out the lesson planning form to be eligible for the paid bonus. You will fill out a lesson plan for each day and each subject you teach. You will fill out the lesson planning form in addition to your MINEDUC lesson journal. Later in this session, you will have a chance to practice using the lesson planning form. You will also see examples of strong and weak lesson plans to help you understand our expectations.

Pedagogy: Teacher observation score

The third component that will affect your eligibility for the paid bonus is your observation score. I will observe your classrooms during the next few weeks at least once, and again next term. I will score your lesson in comparison to other teachers in your district using a rubric. During the observation, I will record all the activities and teaching strategies you use in your lesson. At the end of your lesson, I will use my notes to evaluate your performance in the following four categories:

- Lesson objective, does your lesson have a clear objective?;
- Teaching activities, does your lesson include activities that will help students learn the lesson?;
- Assessment and exercises, does your lesson include exercises for students to practice the skill?; and
- Student engagement, are students engaged during the lesson and activities?

I will use a scoring rubric designed by IPA, Georgetown, and Oxford University to evaluate your performance in each category. You will receive a score from 0 (unsatisfactory) to 3 (exemplary) in each category. I will observe your entire lesson, from beginning to end. I will then evaluate your performance based on the observation. You will not know when I am coming to observe your lesson, so it is in your best interest to plan your lessons everyday as if I were coming to observe. After the lesson, I will share your results with the Head Teacher. You will be able to obtain a copy of your scores, together with an explanation, from the Head Teacher.

Performance: Student test scores

[Field supervisor now speaks.] Half of your overall evaluation will be determined by the learning achievements of your students. We have devised a system to make sure that all teachers compete on a level playing field. If students in your school are not as well off as students in other schools, you do not have to worry: we are rewarding teachers for how much their students can improve, not for where they start.

Here is how this works. We randomly selected a sample of your students to take a cumulative test, testing their knowledge of grade level content. These tests were designed based on the curriculum, to allow us to measure the learning of students for each subject separately. The performance of each teacher will be measured by the learning outcomes of students in the subjects and streams that they themselves teach. (So, if you are a P4 Maths Teacher, your performance will not be affected by students' scores in P4 English. And if you teach P4 Math for Stream A but not Stream B, your performance measure will not depend on students' scores in Stream B.) We will compare the marks for this test with those from other students in the same district, and place each student into one of ten groups, with Group 1 being the best performing, Group 2 being the next-best performing, and so on, down to Group 10. In the district as a whole, there are equal numbers of students placed in each of these groups, but some of your students may be in the same group, and there may be some groups in which you do not have any students at all.

At the end of this school year, we will return to your school and we will sample 10 new pupils from every stream in Upper Primary school to take a new test. This will be a random sample. We do not know in advance who will be drawn, and students who participated in the initial assessment have the same chance of appearing in the end-of-year sample as anyone else. We will draw students for this assessment based on the student enrollment register. If any student from that register is asked to participate in the test but is no longer enrolled at the school, they will receive a score of zero. So, you should do your best to encourage students to remain enrolled and to participate in the assessment if asked. Once the new sample has taken the assessment, we will sort them into groups, with the best-performing student from the final assessment being placed into the group that was determined by the best-performing student in the initial assessment. The second-best student from the final assessment will be placed into the second-highest group achieved from the initial assessment, and so on, until all students have been placed into groups. We will then compare your students' learning levels with the learning levels of other students in the same group only. Each of your students will receive a rank, with 1 being the best, 2 being the next, and so on, within their group. (This means there will be a 1st-ranked student in Group 1, and another student ranked first in Group 2, and so on.) The measure of your performance that we will use for your score is the average of these within-group ranks of the students whom you teach.

This all means that you do not have to have the highest performing students in the District in order to be ranked well. It is possible to be evaluated very well even if, for example, all of your students are in Group 10, the lowest-performing group: what matters is how they perform relative to other students at the same starting point. I will now demonstrate how this works with some examples. Please feel free to ask questions as we go along.

Worked example 1 [Field supervisor sets up Student Test Scores Poster and uses the Student Test Scores Figures to explain this example step by step.] Let us see how the learning outcomes score works with a first example. For this example, suppose that we were to sample 5 students from your class in both the beginning-of-year and end-of-year assessments. (In reality there will be at least ten, but this is to make the explanation easier.) Now, suppose in the initial assessment, we drew 5 students. And those students' scores on the assessment might mean that they are placed as follows:

- One student in Group 1 (top);
- One student in Group 3;
- One student in Group 6;
- One student in Group 9; and
- One student in Group 10.

Then, at the end of the school year, we will return and we will ask 5 new students to sit for a different assessment. These are unlikely to be the same students as before. Once they have taken the test, we will rank them, and we will put the best-performing of the new students into Group 1, the next-best-performing of the new students into Group 3, the next-best performing of the new students into Group 6, then Group 9, and Group 10. So, the Groups into which the new students are placed are determined by the scores of the original students.

Finally, we will compare the actual scores of the new students to the other new students from schools in this district who have been placed into the same groups. For example:

- The new student placed into Group 1 might be ranked 1st within that group;
- The new student placed into Group 3 might be ranked 7th within that group;
- The new student placed into Group 6 might be ranked 4th within that group;
- The new student placed in Group 9 might also be ranked 4th within her group;
- The new student placed into Group 10 might be ranked 1st within his group.

Then, we add up these ranks to determine your score: in this case, it is $1 + 7 + 4 + 4 + 1 = 17$. That is pretty good! Remember, the lower the sum of these ranks, the better. And notice that even though the student in Group 10 did not have a very high score compared to everyone else in the district, he really helped your performance measure by doing very well within his group.

Worked example 2 Now, let us try a second example. Again let us suppose that we were to sample 5 students from your class in both the beginning-of-year and end-of-year assessments. (Remember: in reality there will be at least ten, but this is to make the explanation easier.) Now, suppose in the initial assessment, we drew 5 students. And those students' scores on the assessment might mean that they are placed as follows:

- One student in Group 1 (top);
- TWO students in Group 3;
- One student in Group 4; and
- One student in Group 5.

Notice that it is possible for two or more of your students to be in the same group. Then, at the end of the school year, we will return and we will ask 5 new students to sit for a different assessment. Again, these are unlikely to be the same students as before. Now, suppose that one out of the five students that we ask for has dropped out of school, or fails to appear for the test. They will still be counted, but their exam will be scored as if they answered zero questions correctly—the worst possible score. Once they have taken the test, we will rank them, and we will put the best-performing of the new students into Group 1, the *two* next-best-performing of the new students into Group 3, the next-best performing of the new students into Group 4. The student who was not present for the test because they had dropped out of school is placed into Group 5. As in the previous example, notice that the groups into which the new students are placed are determined by the scores of the original students.

Finally, we will compare the actual scores of the new students to the other new students from schools in this district who have been placed into the same groups. For example:

- The new student placed into Group 1 might be ranked 1st within that group;
- The new students placed in Group 3 might be ranked 4th & 7th in that group;
- The new student placed into Group 4 might be ranked 8th within that group;

- The new student placed in Group 5, who did not actually take the test, will be placed last in his group. If there are 40 students in the group from across the whole district, then this would mean that his rank in that group is 40th.

Then, we add up these ranks to determine your score: in this case, it is $1+4+7+8+40=60$. Notice three points. First, even though in this example, your students did better on the initial assessment than in the first example, this does not mean that you scored better overall. All groups are counted equally, so that no school or teacher will be disadvantaged in this process. Second, notice that the student who dropped out was ranked worst out of the group to which he was assigned. Since the lowest-performing student in the initial assessment was in Group 5, the student who had dropped out was compared with other students placed into Group 5. Since he received the worst possible score, he was ranked last (in this case, fortieth) within that group. This was bad for the teacher's overall performance rank. Third, teachers will be evaluated based on the same number of students. So even if a teacher would be teaching in several streams, resulting in more students taking the tests, his final score will be based on a random subsample of students, such that all teachers are evaluated on the same number of students.

Online Appendix References

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

Bau, Natalie and Jishnu Das, "Teacher value added in a low-income country," *American Economic Journal: Economic Policy*, 2020, *12* (1), 62–96.

Chamberlain, Gary, "Multivariate regression models for panel data," *Journal of Econometrics*, 1982, *18* (1), 5–46.

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

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

Dohmen, Thomas and Armin Falk, "You Get What You Pay For: Incentives and Selection in the Education System," *Economic Journal*, August 2010, *120* (546), F256–F271.

Eckel, Catherine and Philip Grossman, "Men, Women and Risk Aversion: Experimental Evidence," 2008, *1*, 1061–1073.

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

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

Lang, Frieder R., Dennis John, Oliver Ludtke, Jurgen Schupp, and Gert G. Wagner, "Short assessment of the Big Five: robust across survey methods except telephone interviewing," *Behavior Research Methods*, March 2011, *43*, 548–567.

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

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

Mundlak, Yair, "On the pooling of time series and cross section data," *Econometrica*, 1978, *46* (1), 69–85.

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