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

May 7, 2019

World Bank


WUMVA WABA
MWARIMU
S.T.A.R.S?
RCB
WUMVA WABA
MWARIMU
S.T.A.R.S?

Performance pay in the civil service

The ability to recruit, elicit effort from, and retain civil servants is a central challenge of state capacity in developing countries.

- (Finan et al. 2017)

Accumulating evidence that pay-for-performance contracts can elicit effort from incumbent civil servants, although impacts sensitive to design and complementary inputs.

- (Olken et al. 2016, Muralidharan & Sundararaman 2011, Gilligan et al. 2018, Mbiti et al. 2018)

But less is known about how pay-for-performance contracts affect the composition of the civil service.

- (cf. Ashraf et al. 2016, Dal Bó et al. 2013, Deserranno 2017)

Composition and effort margins of performance pay

Two contrasting views:

1. **Pessimistic** (e.g., Bénabou & Tirole 2003, Delfgaauw & Dur 2007, Francois 2000).
   Pay-for-performance contracts worsen outcomes by...
   - Recruiting the wrong types—individuals who are 'in it for the money';
   - Lowering effort by reducing (intrinsic) motivation;
   - Failing to retain the right types—good individuals become de-motivated and quit.

2. **Optimistic** (Lazear 2000, 2003, Rothstein 2015).
   Pay-for-performance contracts improve outcomes by...
   - Recruiting the right types—individuals who anticipate performing well;
   - Raising effort by increasing (extrinsic) motivation;
   - Retaining the right types—good individuals feel rewarded and stay put.

Teachers matter

With rising access to government schooling failing to translate into hoped-for learning gains in many developing countries. . .

- Teaching salaries account for the bulk of education expenditure (Das et al 2017).
- Teacher value-added has persistent effects on learning and subsequent labor-market outcomes (Chetty et al. 2014*a*,*b*).
- There is substantial variation in teacher value added within a given school system (Buhl-Wiggers et al. 2016).
- Teachers' mastery of the curriculum is in many places a challenge: for example, World Bank's SDI estimates that only 20 percent of Ugandan fourth-grade teachers have mastery of grade-level content.

Screening for teacher quality is hard—if not impossible

Teacher quality, measure by value added, is difficult to predict using pre-employment characteristics (Hanushek & Rivkin 2006, Rockoff et al. 2011).

But: theory suggests the extensive-margin effects of performance contracts may be substantial (Lazear 2000, 2003, Rothstein 2015).

While an emergent body of literature suggests P4P may deliver learning gains for students of current teachers, no rigorous evidence of its compositional consequences in developing countries.

Performance contracts and teacher quality

Hanushek on the teacher quality equilibrium in the U.S.

> *"If we just raise all teacher salaries, we are going to raise the salaries of current effective teachers and current ineffective teachers, and we are going to lock in our current workforce for a while into the future because it's an attractive job, and more attractive with higher pay.*
>
> *So the only answer from a policy standpoint if we want to change our achievement within the next two decades is to think of a bargain, where we increase the pay of teachers, but also—at the same time—tilt the function more based on the effectiveness of teachers."*

This study provides the first prospective, experimental evidence of P4P effects on civil-service composition, effort, and retention.

Project genesis and timeline

- PIs commissioned to write a white paper on policies for education quality—including teacher management—by SPU as an input into the National Leadership Retreat 2014.
- Pilot program designed for the 2015 school year, in consultation with a REB stakeholder/advisory group.
- Materials developed for assessment of students and teachers shared with REB.
- Results submitted to GoR in early 2016 and presented in person to then-DG REB.
- Phase II districts identified with REB in July 2015
- Workshop with Phase II districts in September 2015
- Recruitment of teachers undertaken into 2016 school year.
- Project implemented in 2016 and 2017 school years.
- Blinded data used to develop specifications and analyses.

Policy context and fit

Study is aligned with several features of the education sector and civil service:

- *Imihigo* in other sectors provides a framework. Study performance contracts designed to match typical *imihigo* stakes (3 percent of salary).
- Existing, complementary policy mix seeks to make teaching more attractive, and to reward effective teachers (cows, laptops, *Umwalimu* Sacco).
- Evidence of impacts of performance contracts in health sector.
- National Leadership Retreat 2019 resolution:

  > *7. Strengthen programs to improve the quality of education focusing on. . . [among others] recruitment of more qualified teachers for primary and secondary schools.*

Design

Contracts

Fixed Wage:

An end-of-year payout of RWF 20,000.

Roughly 3 percent of typical wages, on par with typical salary increments and variable pay under the *imihigo* system for the rest of the civil service.

Pay-for-Performance:

An end-of-year payout of RWF 100,000 for those in the top quintile, or zero otherwise.

Performance metric puts 50% weight each on:

- Learning outcomes: Barlevy and Neal 2012—average end-of-year rank of students within bands defined by baseline outcomes.
- Teachers' effort: preparation, presence, pedagogy.

Study design: Two-stage randomization

Study working in 6 districts:
Gatsibo, Kayonza, Kirehe, Ngoma, Nyagatare, Rwamagana.

Potential applicants in each divided by subject of qualification:
Modern Languages, Math & Science, Social Studies.

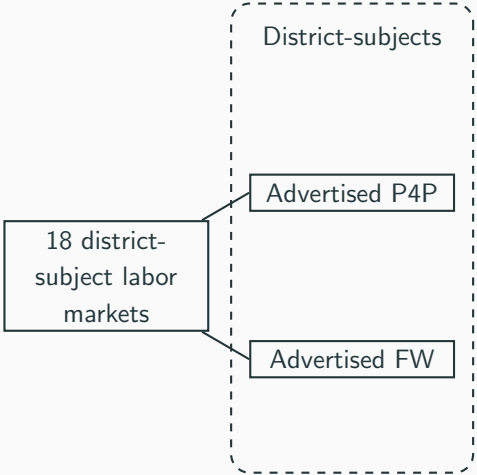
Resulting 18 'labor markets', comprise 600+ hiring lines, and more than 60% of planned hiring in 2016.

18 district-subject labor markets

Study design: Two-stage randomization

Labor markets are randomly assigned to Advertised P4P or Advertised FW (or Advertised Mixed, not represented here.)

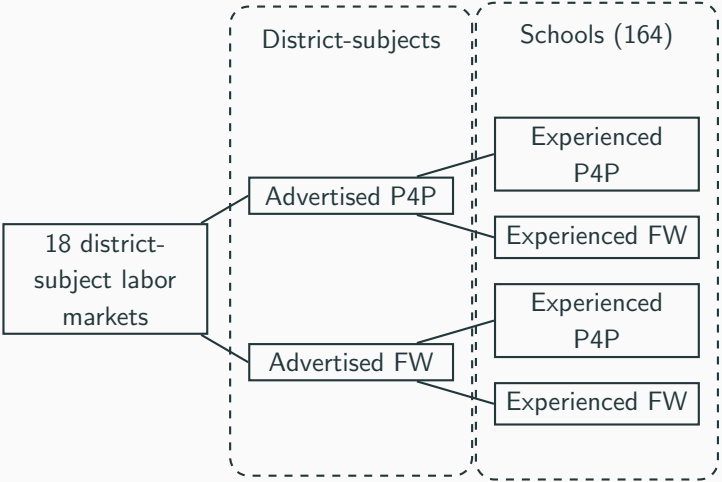
Comparison of applicant and hired-teacher characteristics across these markets reveals the effect of advertisement.



Study design: Two-stage randomization

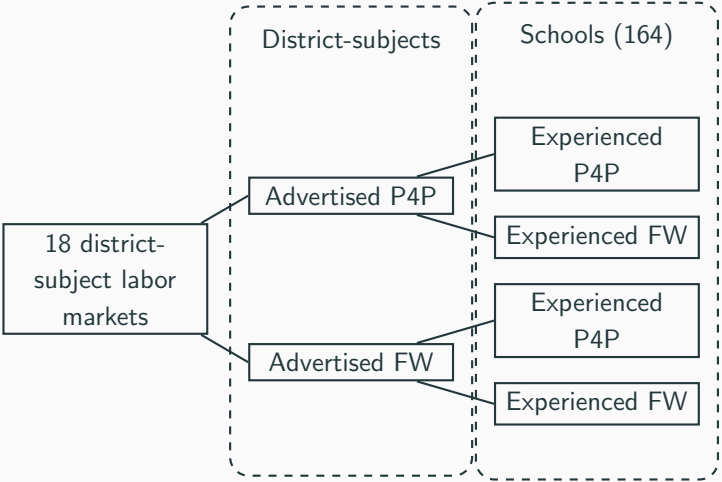
Hired teachers are placed in upper-primary positions in 164 schools.

Schools randomly assigned to Experienced P4P or FW.



Study design: Two-stage randomization

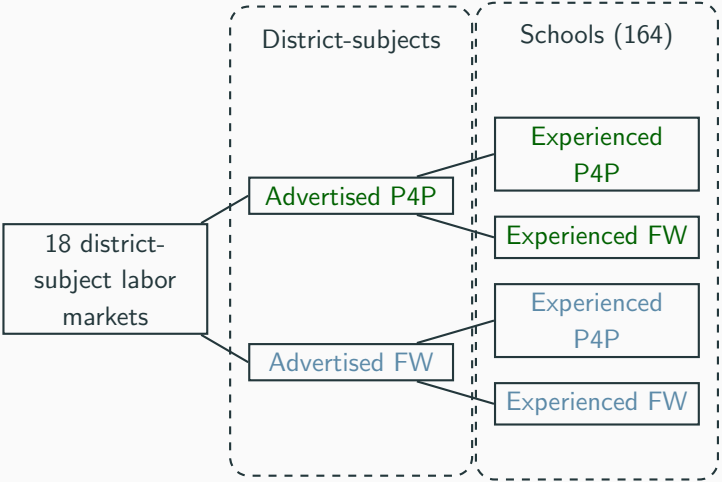
This design enables three comparisons:



Study design: Two-stage randomization

This design enables three comparisons:

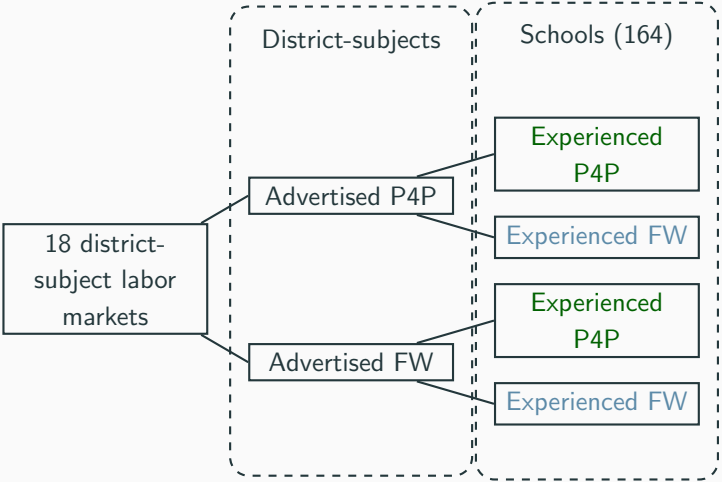
- Advertised P4P vs Advertised FW reveals *recruitment effect*.



Study design: Two-stage randomization

This design enables three comparisons:

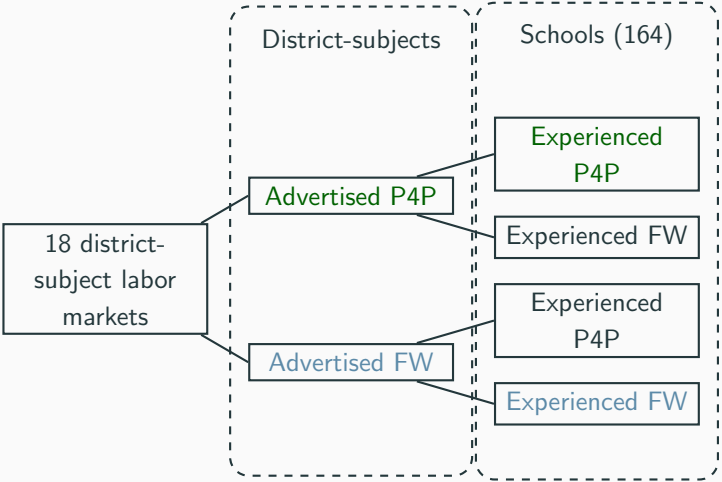
- Advertised P4P vs Advertised FW reveals *recruitment effect*.
- Experienced P4P vs Experienced FW reveals *effort response*.



Study design: Two-stage randomization

This design enables three comparisons:

- Advertised P4P vs Advertised FW reveals *recruitment effect*.
- Experienced P4P vs Experienced FW reveals *effort response*.
- Advertised + Experienced P4P vs Advertised + Experienced FW reveals *total effect*.



Outcomes

1. Applications. We observe the universe of applications in study districts. TTC exam scores, gender, district application exams.
2. Placed teacher characteristics. For teachers in upper-primary posts, measure skills, motivation, and a battery other characteristics at baseline.
3. Learning. Learning gains over the year in grade-stream-subjects taught by recruits.
4. Teacher inputs. Contracted measures of presence, preparation (lesson plans), and pedagogy (Danielson-based classroom observation score). P4P schools year 1; all schools year 2.

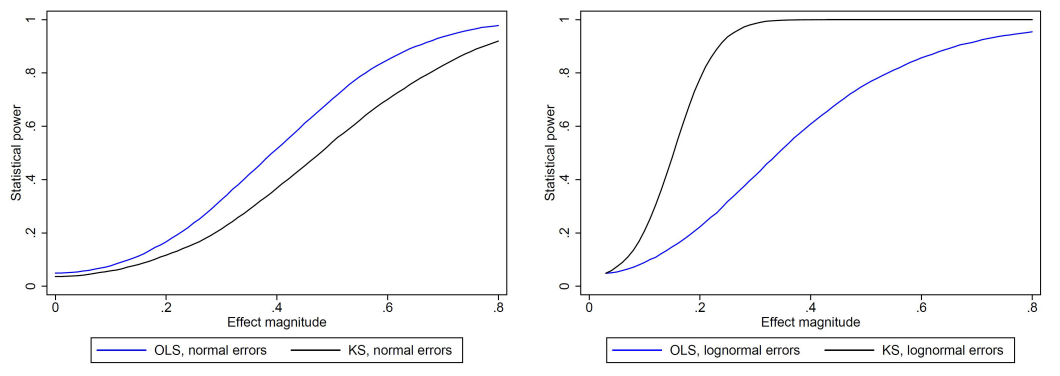
Analysis plan

In our pre-analysis plan, we address the question of *how* to provide well-powered tests of hypotheses using blinded data.

For example:

- Kolmogorov-Smirnov test vastly outpowers regression-based tests of changes in application characteristics, even against additive shifts.
- Linear mixed-effects model (with pupil-round random effects) using data from incumbents' pupils minimizes standard deviation of recruitment and effort-margin effects under the 'sharp' null.

Example: OLS is more powerful with normally distributed errors; KS with log-normal



Simulated power for TTC scores in application pool

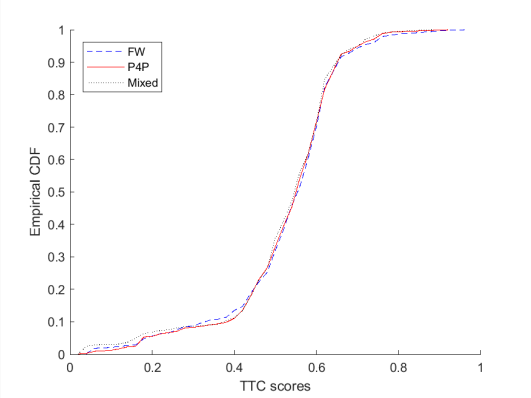
Simulated rejection rates for treatment effects that move a candidate at the median by 1, 2, 5, or 10 percentile ranks:

Test statistic	τ_1	τ_2	τ_3	τ_4
T^{KS}	0.45	1.00	1.00	1.00
T^{OLS}	0.11	0.37	0.92	1.00

Results

Formal qualifications in the applicant pool

No impact of advertised P4P on Teacher Training College final exam score among applicants (Hypothesis I).

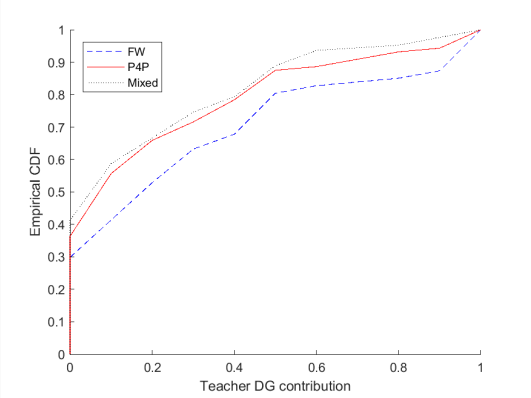


RI is well-powered, can rule out even small positive effects on TTC score distribution.

Likewise, no impact on our baseline assessment of skill among placed recruits (Hypothesis II).

Teacher motivation on arrival

Advertised P4P did impact our baseline assessment of 'intrinsic motivation' among placed recruits (Hypothesis III).



We asked recruits to divide a small amount of money between themselves and support for students in their placement school.

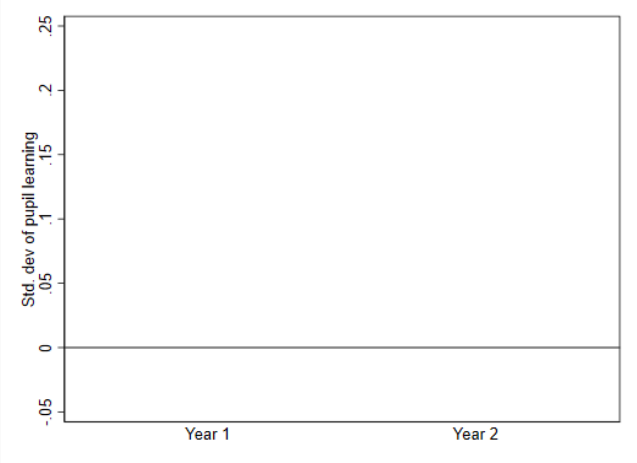
Teachers recruited under advertised FW contracts were significantly more generous.

Effects on student learning

We estimate a linear mixed-effects model of form

$$
\begin{aligned}
z_{jbksr} \quad = \quad & \tau_A T^A_{qd} + \tau_E T^E_s \\
& + \lambda_I I_i + \lambda_E T^E_s I_i \\
& + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e
\end{aligned}
$$

where T^A_{qd} is Advertised P4P for qualification q in district d, T^E_s is Experienced P4P in school s, I_i indicates that the teacher is an incumbent, $\bar{z}_{ks,r-1}$ are lagged mean test scores, and δ_d and ψ_r are district and round fixed effects.



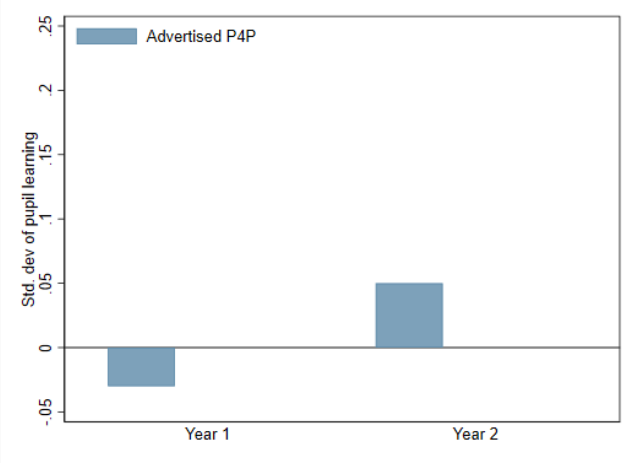
Learning impacts table

Effects on student learning

We estimate a linear mixed-effects model of form

$$
\begin{aligned}
z_{jbksr} \quad = \quad & \tau_A T^A_{qd} + \tau_E T^E_s \\
& + \lambda_I I_i + \lambda_E T^E_s I_i \\
& + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e
\end{aligned}
$$

where T^A_{qd} is Advertised P4P for qualification q in district d, T^E_s is Experienced P4P in school s, I_i indicates that the teacher is an incumbent, $\bar{z}_{ks,r-1}$ are lagged mean test scores, and δ_d and ψ_r are district and round fixed effects.



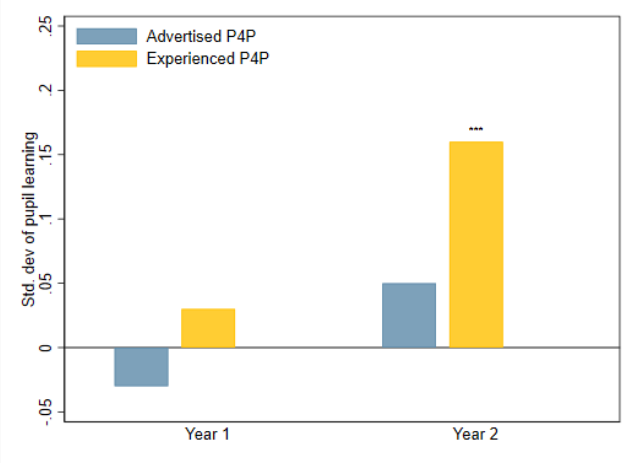
▸ Learning impacts table

Effects on student learning

We estimate a linear mixed-effects model of form

$$
\begin{aligned}
z_{jbksr} \quad = \quad & \tau_A T^A_{qd} + \tau_E T^E_s \\
& + \lambda_I I_i + \lambda_E T^E_s I_i \\
& + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e
\end{aligned}
$$

where T^A_{qd} is Advertised P4P for qualification q in district d, T^E_s is Experienced P4P in school s, I_i indicates that the teacher is an incumbent, $\bar{z}_{ks,r-1}$ are lagged mean test scores, and δ_d and ψ_r are district and round fixed effects.



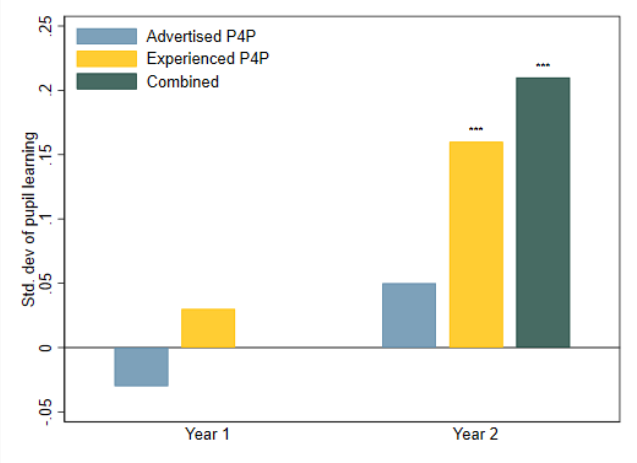
▸ Learning impacts table

Effects on student learning

We estimate a linear mixed-effects model of form

$$
\begin{aligned}
z_{jbksr} \quad = \quad & \tau_A T^A_{qd} + \tau_E T^E_s \\
& + \lambda_I I_i + \lambda_E T^E_s I_i \\
& + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e
\end{aligned}
$$

where T^A_{qd} is Advertised P4P for qualification q in district d, T^E_s is Experienced P4P in school s, I_i indicates that the teacher is an incumbent, $\bar{z}_{ks,r-1}$ are lagged mean test scores, and δ_d and ψ_r are district and round fixed effects.



▸ Learning impacts table

Impacts on the performance metric

We also find significant impacts of experienced P4P on the incentivized composite performance metric (Hypothesis VI).

Secondary analysis suggests that learning outcomes improved because the pay-for-performance contracts elicited more effort on two dimensions.

1. Presence. 6 percentage points higher among recruits who experienced the P4P contract compared to recruits who experienced the FW contract.
2. Pedagogy (as measured on a four-point classroom practice scale). 0.26 points higher among recruits who experienced the P4P contract compared to recruits who experienced the FW contract. N.B. 21 activities observed over 45 minutes.

Inputs table

Are (growing) P4P effects due to exit?

We find no evidence of differential attrition by Experienced P4P, or its interaction with baseline skill and motivation.

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

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

Our findings so far

1. Pay-for-performance contracts changed the composition of the teaching workforce, drawing in individuals who were more money-oriented.
    - *But these recruits were not less effective teachers, if anything the reverse.*
2. Pay-for-peformance contracts raised teacher effort, notably in terms of presence and pedagogy.
3. No evidence that pay-for-performance contracts impacted retention.

These effects combined to *raise learning quality*.

Policy takeaways

A P4P model that has potential to improve learning outcomes.

- Modest impact in Year 2, if anything stronger net of recruitment margin, and effects may accumulate over years of exposure.
- Potentially budget neutral given existing annual salary increments and plans for annual testing.
- Echoes results in healthcare, and fits with the prevailing *imihigo* system.
- Popular with teachers, easing concerns over implementation.

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

May 7, 2019

World Bank

References i

Ashraf, N., Bandiera, O. & Lee, S. S. (2016), 'Do-gooders and go-getters: Selection and performance in public service delivery', Working paper.

Barlevy, G. & Neal, D. (2012), 'Pay for percentile', *American Economic Review* **102**(5), 1805–1831.

Bénabou, R. & Tirole, J. (2003), 'Intrinsic and extrinsic motivation', *Review of Economic Studies* **70**, 489–520.

Buhl-Wiggers, J., Kerwin, J. T., Smith, J. A. & Thornton, R. (2016), 'The impact of teacher effectiveness on student learning in Africa', Unpublished.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014*a*), 'Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates', *American Economic Review* .

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014*b*), 'Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood', *American Economic Review* **104**(9), 2633–2679.

Dal Bó, E., Finan, F. & Rossi, M. (2013), 'Strengthening state capabilities: The role of financial incentives in the call to public service', *Quarterly Journal of Economics* **128**(3), 1169–1218.

Delfgaauw, J. & Dur, R. (2007), 'Incentives and workers' motivation in the public sector', *Economic Journal* **118**(525), 171–191.

References ii

Deserranno, E. (2017), 'Financial incentives as ssignal: Experimental evidence from the recruitment of village promoters in Uganda', Working paper.

Finan, F., Olken, B. A. & Pande, R. (2017), The personnel economics of the state, *in* 'Handbook of Field Experiments', Vol. 2, Elsevier, pp. 467–514.

Francois, P. (2000), "public service motivation' as an argument for government provision', *Journal of Public Economics* **78**(3), 423–464.

Gilligan, D., Karachiwalla, N., Kasirye, I., Lucas, A. & Neal, D. (2018), 'Educator incentives and educational triage in rural primary schools', NBER Working Paper No. 24911.

Hanushek, E. A. & Rivkin, S. G. (2006), Teacher quality, *in* E. Hanushek & F. Welch, eds, 'Handbook of the Economics of Education', Vol. 2, Elselvier B. V., Amsterdam, pp. 1051–1078.

Lazear, E. P. (2000), 'Performance pay and productivity', *American Economic Review* **90**(5), 1346–1361.

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

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Rajani, R. & Manda, C. (2018), 'Inputs, incentives, and complementarities in primary education: Experimental evidence from Tanzania', Working paper, University of Virginia.

References iii

Muralidharan, K. & Sundararaman, V. (2011), 'Teacher performance pay: Experimental evidence from India', *Journal of Political Economy* **119**(1), 39–77.

Olken, B. A., Khan, A. & Khwaja, A. (2016), 'Tax farming redux: Experimental evidence on performance pay for tax collectors', *Quarterly Journal of Economics* **131**(1), 219–271.

Rockoff, J. E., Jacob, B. A., Kane, T. J. & Staiger, D. O. (2011), 'Can you recognize an effective teacher when you hire one?', *Education Finance and Policy* **6**(1), 43–74.

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

Supplemental results

Blinded pre-analysis plan: Primary hypotheses, measures, and specifications

Hypothesis I. Advertised P4P induces differential application qualities;

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

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

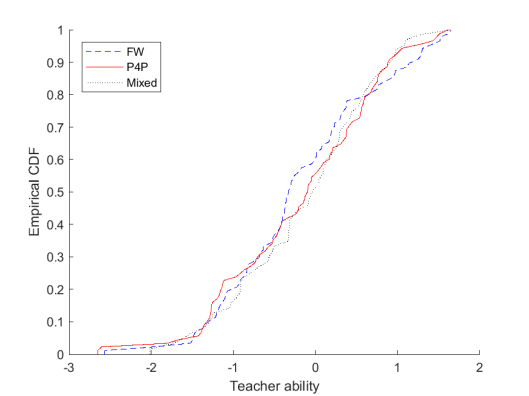
Hypothesis IV. Advertised P4P induces the *selection* of higher- (or lower-) performing teachers, as measured by the learning outcomes of their students;

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

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

Consistent with application data, we find no evidence of selection on skill among placed recruits. . .

Estimated impact of P4P recruitment on recruit ability is -0.17 ($p = 0.46$).



Other attributes of hired teachers under P4P

	τ_A	CI	p	N
lottery choice	-0.11	[-0.60, 0.37]	0.62	238
tournament choice	-0.12	[-0.27, 0.03]	0.08	235
big5std	-0.03	[-0.24, 0.19]	0.79	239
locusoc	-0.11	[-0.31, 0.08]	0.17	236
selfesteem	-0.32	[-1.35, 0.74]	0.47	236
age	0.05	[-0.91, 1.01]	0.92	312
female	0.15	[-0.04, 0.34]	0.09	281

Impacts of advertised and experienced contracts on TVA, by year

We estimate a linear mixed-effects model of form

$$
\begin{aligned}
z \quad = \quad & \tau_A T^A_{qd} + \tau_E T^E_s \\
& + \lambda_I I_i + \lambda_E T^E_s I_i \\
& + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e
\end{aligned}
$$

where T^A_{qd} is advertised P4P for qualification q in district d, T^E_s is experienced P4P in school s, I_i indicates that the teacher is an incumbent, $\bar{z}_{ks,r-1}$ are lagged mean test scores, and δ_d and ψ_r are district and round fixed effects.

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

Randomization inference p-values in brackets.

Back

Impacts of experienced contracts on teacher inputs

From a (school-year random effects) model of the form

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

we estimate:

	BN rank: round 2	Presence: round 2	Pedagogy: round 2
τ_A	0.08	0.01	0.11
	[0.43]	[0.48]	[0.42]
τ_E	0.10	0.06	0.26
	[0.07]	[0.08]	[0.07]
λ_E	0.04	-0.05	0.07
	[1.00]	[0.98]	[0.98]

Randomization inference p-values in brackets.

Back