Do performance contracts retain better teachers, and do they (de)motivate? A blinded pre-analysis plan

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

September 27, 2018

Contents

Appendices

1 Introduction

This pre-analysis plan forms part of a wider project, and should be read alongside the companion plan (Leaver, Ozier, Serneels and Zeitlin, 2018a). In that pre-analysis plan, we exploit the first-stage randomization of new teaching posts in distinct labour markets to either advertised pay-for-performance (P4P) contracts or advertised fixed wage (FW) contracts. Our focus there is exclusively on treatment impacts upon applicants and new recruits,[1] with the objective of establishing whether advertised P4P has a significant recruitment effect. That is, do more individuals apply, and are they of better quality, as measured by their baseline skill, intrinsic motivation, and subsequent on-the-job performance (teacher value-add to student learning)?

In this plan, we switch focus from the compositional effect of *advertised* P4P at the recruitment stage to the impact of *experienced* P4P. To do so, we exploit the second-stage randomization of schools to either experienced P4P or FW contracts. For clarity, we focus exclusively on treatment impacts upon incumbents who were not subject to the first-stage randomization.[2] We develop a simple theoretical framework (set out in Appendix A) which distinguishes between three forces: extrinsic incentives, intrinsic motivational crowding out, and retention. Guided by this framework, our objective in this second plan is to quantify the long-term impact of P4P and to decompose any such impact into its constituent parts. Specifically, the three primary research questions are:

A. Does experienced P4P result in better teacher performance than experienced FW?

B. Teacher retention: Are better teachers retained across years under experienced P4P than experienced FW?

C. Teacher (de)motivation: Does experienced P4P impact the intrinsic motivation of retained teachers?

As secondary research questions, we are also interested in heterogeneous treatment effects by school characteristics (especially management, as in Leaver, Lemos and Scur, 2018b), baseline teacher characteristics (risk aversion, taste for competition, socioemotional traits), and student characteristics.

2 Study design

A full description of the study design can be found in Leaver et al. (2018a).

3 Measurement and data

A full description of our measurement activities and the data we collected can be in found in Leaver et al. (2018a). See in particular the study profile in Figure 2.

In addition to the data described there, we collected several repeat measures of teacher motivation at the Year 2 endline to allow testing of Hypothesis C. These include the incentivized decisions in the following lab-in-the-field measures: Dictator Game allocations, decisions in the Lottery Choice Game, and decisions in the Contract Choice Game. They also include survey measures of job satisfaction, positive and negative affect, the Perry public-service motivation instrument, the Big Five Index, and the Locus of Control.

[1] We use data on incumbents purely for power purposes (to estimate random effects).

[2] We will report results for the small subset of new recruits in an Appendix. The theory covers both.

4 Empirical specifications and inference

We set out to test the three specific questions referred to in the Introduction. We , resulting in following specific questions. Each of these has a primary hypothesis and a small number of associated secondary hypotheses that represent alternative measures or mechanisms. These hypotheses are:

A. Does experienced P4P result in better teacher performance than experienced FW?

   I. Experienced P4P affects the value-added of incumbents;

   II. Experienced P4P affects the performance of incumbents on the composite 4P metric.

B. Are better teachers retained across years under experienced P4P than experienced FW?

   III. Experienced P4P affects retention rates among incumbents;

   IV. Experienced P4P induces differentially skilled incumbents to be retained;

   V. Experienced P4P induces differentially 'intrinsically' motivated incumbents to be retained;

   VI. Experienced P4P induces better (Year 1 TVA-ranked) incumbents to be retained.

C. Does experienced P4P impact the intrinsic motivation of retained teachers?

   VII. Experienced P4P changes within-retained-incumbent intrinsic motivation from baseline to endline by more (or less).

We detail outcome measures and specifications for each of these hypotheses below.

4.1 Hypothesis I: Experienced P4P affects the value-added of incumbents

In Leaver et al. (2018a), we explore whether advertised P4P affects the value-added of recruits, through both compositional and effort margins. In doing so, we identify a pure selection effect on the recruitment margin, abstracting from any incentive effect. By contrast, in this second pre-analysis plan, we are interested in the overall performance effect, and the extent to which this is driven by incentives versus selection on the *retention* margin.

Our primary specification focuses on the overall impact of the second stage school-level randomization on student learning. As in our first paper, the measure of student learning which we deploy is the Empirical Bayes prediction of student ability, based on the IRT model of student assessments. This is observed at the student-subject level; each sampled student takes an assessment in all five core subjects. We follow the same notation as before and denote by z_{jbkgsr} this measure of learning for student j in subject b, appearing in stream k, school s, and round r.

Our primary estimand is the impact of the school-level contract on incumbent teachers' annual value-added. We will pool data across the two years of intervention and estimate as primary a specification of the type

$$z_{jbkgsr} = \tau_E T_s^E + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e_{jbkgsr}. \tag{1}$$

Here, δ_d and ψ_r represent coefficients on indicators for districts and rounds, respectively. In all specifications, we will include a vector of stream-mean lagged test scores $\bar{z}_{ks,r-1}$.[3] When the

[3] We use stream k mean lagged assessment scores, rather than lags specific to student j, because the rotating sample of students within a given stream means that a restriction to students for whom consecutive test scores are available would have a restrictive effect on sample size, and power. Representative sampling in each round implies that this specification should be equally free of bias.

outcome is year-one endline assessments, $r = 1$, these lagged assessment scores correspond to baseline scores, but when the outcome is year-two assessments, lagged test scores correspond to year-one outcomes.[4] The association between lagged and current assessments are allowed to vary by subject, grade, and round, as indicated by the parameter ρ_{bgr}.

As in our analysis of recruitment effects (Leaver et al., 2018a), we estimate equation (1) as a linear mixed effects model, allowing for unobserved, normally distributed heterogeneity at the student (j) level. This model is estimated by maximum likelihood. We employ this LME model as our primary approach to estimating all subsequent empirical specifications that use student assessments as the outcome.

In addition to this primary specification, we will obtain secondary estimates of the impact of the experienced P4P treatment for year-one outcomes alone and year-two outcomes alone. Across these primary and secondary estimates, our object of interest is the parameter, τ_E. In the pooled primary specification and in the secondary year-two alone specification, τ_E should be interpreted as the overall performance effect (i.e. combining any extrinsic incentive, intrinsic motivational crowding, and retention effects). When the outcome is year-one outcomes alone, and under the assumption that intrinsic motivational crowding evolves slowly, τ_E plausibly isolates any extrinsic incentive effect.

We will obtain an estimate of the distribution of τ_E under the sharp null hypothesis of no treatment effect for any unit by randomization inference permuting the school-level treatment. As in Leaver et al. (2018a), we studentize this parameter by dividing it by its estimated standard error, and use the studentized parameter, t_E, as as the relevant test statistic in this permutation test (see DiCiccio and Romano, 2017; Chung and Romano, 2013, for details).[5]

As part of our secondary analysis, we will also explore the extent to which individual teachers' response to experienced P4P depends on their own, baseline attributes. To do so, we will use a specification of the form

$$z_{jkbsr} = \tau_E T_s^E + \zeta T_s^E X_{is0} + \beta X_{is0} + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e_{jkbsr}, \tag{2}$$

where X_{is0} are baseline characteristics of the teacher i assigned to subject-stream-grade combination kbg, such that $i = i(kbg)$. Specifically, we test individually and jointly for heterogeneity in response by the following teacher characteristics: baseline Dictator Game allocation, Grading Task test score, risk aversion, taste for competition, and personality (Big Five index). Within this family of teacher characteristics and in the subsequent families of heterogeneity below, we follow (Young, 2017) to use the minimum p value across these dimensions of heterogeneity as a test for joint significance of the associated interaction terms ζ, in a randomization inference context.

A further secondary analysis will examine heterogeneity in impacts by students' baseline abilities. To do so, we will use the subset of panel students in our dataset and interact the experienced P4P treatment, T_s^E, with the student's own baseline ability in the same subject, $z_{jkbs,r-1}$, in an analogous modification of the specification in equation (1).

A final part of the secondary analysis will investigate impact heterogeneity in terms of school level characteristics, including management practices, socioeconomic profile the school's students, school facilities and Head Teacher skills and characteristics.

[4] See our first paper for a discussion.

[5] Because the LME model is estimated by maximum likelihood, technically, this estimate is a z statistic rather than a t statistic, but in any case we are not comparing the resulting statistic to a reference distribution, normal or otherwise, but only to its distribution under the sharp null as produced by randomization inference.

4.2 Hypothesis II: Experienced P4P affects the performance of incumbents on the composite 4P metric

Experienced P4P contracts reward teachers on the basis of a composite performance metric based on both teacher inputs (presence, preparation, and pedagogy) and a Barlevy and Neal-style measure of student learning outcomes (Barlevy and Neal, 2012).[6] To observe teachers' more proximate responses to the P4P scheme, we propose to directly estimate impacts of experienced P4P on this composite performance metric.

The sample for this estimate is complicated by the fact that we only observe teacher inputs in P4P schools in Year 1 of the study. Our primary test will therefore necessarily use outcomes from Year 2. As a robustness check, will also estimate this specification (without the round indicator) on Year 2 data only. For the available sample of incumbents in a given year, we then estimate a regression of the form

$$m_{iqsdr} = \tau_E T_s^E + \psi_r + \phi_b + \delta_d + e_{iqsdr} \tag{3}$$

where m_{iqsdr} is the outcome of teacher i, teaching (a vector of) subjects b, in school s of district d, as observed in post-treatment round $r \in \{1, 2\}$. As in Leaver et al. (2018a), we estimate equation (3) with a school-year random effects model, and studentize the coefficient τ_E to conduct hypothesis tests by randomization inference. As secondary specifications we will repeat the analysis for each of the four components of the composite performance metric separately.

The theoretical framework set out in the Appendix predicts that P4P contracts have a retention effect—that is, an impact upon the outflow of incumbents between Years 1 and 2. We investigate this in hypotheses III-VI set out below.

4.3 Hypothesis III: Experienced P4P affects retention rates among incumbents

We will first to test for impacts on the likelihood that an incumbent is still employed at midline in February 2017 at the start of the Year 2 (i.e. after experiencing P4P in Year 1, although before the performance awards were announced). Our primary test of this hypothesis is a linear probability model of the form

$$\Pr[employed_{iqd2} = 1] = \tau_E T_s^2 + \phi_b + \delta_d, \tag{4}$$

where $employed_{iqd2}$ is an indicator variable taking a value of one if teacher i teaching subjects b in district d is still employed by the school at the end of Year 2.

As a secondary specification, we will repeat this analysis for the endline teacher listing. Since this is after Year 1 teacher performance awards were announced in July 2017, we will also explore whether the receipt of a performance award itself has an impact on teacher retention among teacher in the P4P arm (not in the theory but an intuitive possibility). Whether a teacher receives a performance award is determined solely by a threshold in the performance score, so this analysis will be done with a regression discontinuity in the teacher's performance score. We anticipate following the optimal bandwidth and local linear specification of Imbens and Kalyanaraman (2012), checking the robustness of this result by using both the Gelman and Imbens (2018) second-order polynomial suggestion, as well as the optimal bandwidth of Calonico, Cattaneo, and Titiunik (various years).

[6] Appendix B of Leaver et al. (2018a) describes in detail how this composite metric is constructed. We use contractual teacher scores in the P4P arm, while in the FW arm, we impute the rankings that teachers in these schools would have received, based on their observation scores.

4.4 Hypotheis IV: Experienced P4P induces differentially skilled incumbents to be retained

We will also test for impacts on the *quality* of incumbents who are retained, beginning with teacher skill. Our primary measure of teacher skill is the Grading Task administered to all incumbents whose assignment at baseline included at least one upper-primary subject. We will use IRT estimates of teacher ability in this subject, which we denote by z_{iqsd} for teacher i, whose skill was assed in subject q, teaching subjects b in school s and district d. We then estimate impacts on average retained incumbent skill levels using a regression of the form

$$\Pr[employed_{iqd2} = 1] = \tau_E T_s^E + \beta z_{iqsd} + \zeta T_s^E z_{iqsd} + \gamma_q + \phi_b + \delta_d, \tag{5}$$

where γ_q denotes coefficients on a vector of subject-of-teacher-assessment indicators, and ϕ_b and δ_d again denote coefficients on a vector of subject-taught and district indicators, respectively. Inference for the key parameter, ζ, is undertaken by performing randomization inference for alternative assignments of the school-level treatment indicator. (A more general model, allowing the parameter ζ to vary by subject, q, of teacher assessment, will be estimated as secondary, to allow for the possibility that some subject-tests have greater predictive power over teacher retention than others.)

As a secondary test of this hypothesis, we will also examine TTC score variation among retained incumbents as our measure of teacher skill, in lieu of our own teacher skill assessments in equation (5). This will allow us to look at a measure of quality among the full set of teachers hired into a school, since this information is available regardless of whether a given teacher was teaching an upper primary class.

4.5 Hypothesis V: Experienced P4P induces differentially 'intrinsically' motivated incumbents to be retained

In addition to the possibility that experienced P4P may (de)select teachers on the basis of skill, such contracts may also change the distribution of the intrinsic motivation among retained incumbents. To test for such effects, we use a specification analogous to equation (5), again, conducting inference for the sharp null of no effect using randomization inference on the school-level treatment indicator. Specifically, we estimate a linear probability model of the form

$$\Pr[employed_{isd2} = 1] = \tau_E T_s^E + \beta x_{isd} + \zeta T_s^E x_{isd} + \phi_b + \delta_d \tag{6}$$

where x_{isd} is the share of the stake allocated by a retained incumbent teacher to the school in the *baseline* Dictator Game, and all other variables are defined as in equation (5). In secondary specifications, we will test for impacts on retained incumbents' taste for competition, and for their (over)confidence in their own abilities, and their degree of risk aversion. All of these measures are described in detail in Leaver et al. (2018a). As in Hypothesis I, we follow Young (2017) to use the minimum p value across these secondary dimensions of heterogeneity as the randomization inference test statistic in a test for joint significance.

4.6 Hypothesis VI: Experienced P4P induces better (Year 1 TVA-ranked) incumbents to be retained

In the theory set out in the Appendix, teacher skill and intrinsic motivation are two dimensions of teacher 'quality', which alongside effort, contribute to teacher performance as measured by value-added to student learning outcomes. With this in mind, we will also look for evidence of teacher selection by exploring Year 1 teacher value-added *rank* among retained incumbents. (We focus on ranks rather than the absolute levels to abstract from Year 1 incentive effects; ranks will be defined within treatment arm for this purpose.) Our specification is again a variant of (6).

4.7 Hypothesis VII: Experienced P4P changes within-retained-incumbent intrinsic motivation from baseline to endline by more (or less)

Aside from the extrinsic incentive effect, the theory set out in the Appendix suggests that P4P may also impact on the intrinsic motivation of a *given teacher*.[7] We will explore this by estimating the following specification

$$x_{isd2} = \tau_E T_s^E + \rho x_{isd0} + \phi_b + \delta_d + e_{isd} \tag{7}$$

where x_{is2} is the endline (round 2) outcome, x_{isd0} is the baseline (round 0) outcome, ϕ_b is a vector of subject indicators, and δ_d is a vector of district indicators.

[7] We emphasize given teacher because this is conceptually distinct from affecting which baseline teacher types choose to remain in Year 2.

References

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

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

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

Gelman, Andrew and Guido Imbens, “Why high-order polynomials should not be used in regression discontinuity designs,” *Journal of Business & Economic Statistics*, 2018, pp. 1–10.

Imbens, Guido and Karthik Kalyanaraman, “Optimal bandwidth choice for the regression discontinuity estimator,” *Review of Economic Studies*, 2012, *79* (3), 933–959.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin, “Do performance contracts attract better teachers? A blinded pre-analysis plan,” Unpublished, University of Oxford 2018.

_ , Renata Lemos, and Daniela Scur, “Why does management matter? A theoretical framework,” Unpublished, University of Oxford 2018.

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

Appendix A Theory

Appendix A.1 Model

Preferences Individuals care about their compensation w and their effort e. Assume all individuals are risk neutral. The *per-period* payoff is

$$u(w, e, \tau, \text{sector}) = w - c(e, \tau, \text{sector}),$$

where c is a convex cost of effort function, τ denotes 'motivational type', and sector is either 'education' or 'other'. When an individual is working in the education sector this function is a quadratic of the form

$$\begin{aligned} c(e, \tau, \text{education}) &= e^2 - \tau \cdot e \\ \Rightarrow \frac{\partial c}{\partial e} &= 2e - \tau \lesseqgtr 0 \\ \frac{\partial c^2}{\partial^2 e} &= 2 > 0. \end{aligned}$$

So for effort levels $e < \tau/2$, incumbents derive a marginal *benefit* from exerting an extra unit of effort in teaching; it is only when $e > \tau/2$ that effort costs kick in. In this sense, τ captures 'intrinsic motivation to teach'. An incumbent's type T is a uniform random variable with support $[0, \tau^{max}]$. Incumbents observe their draw τ perfectly. Motivational type plays no role in the other sector: $c(e, \tau, \text{other}) = e^2$.

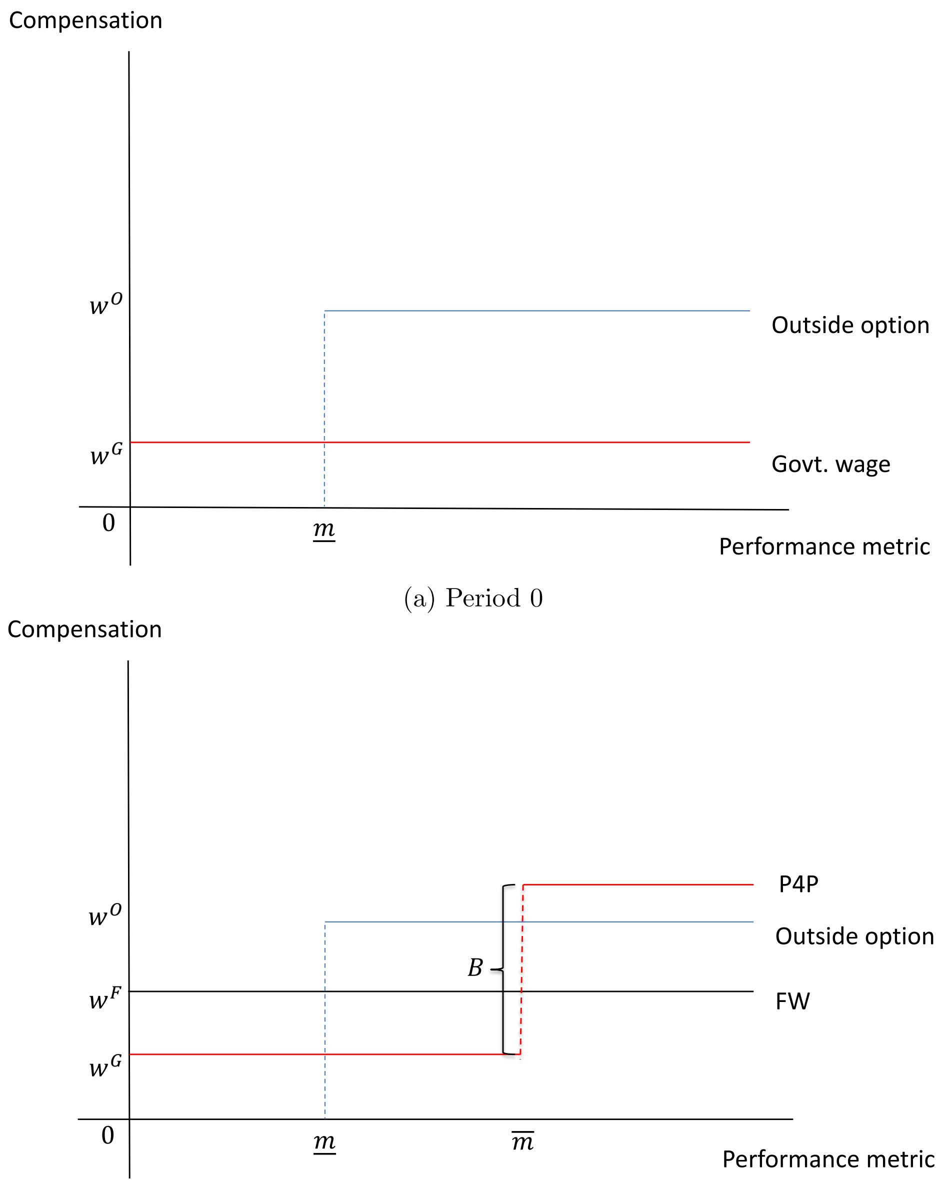
Performance Metric and Compensation Schemes Irrespective of where an individual works, each period her effort generates a performance metric

$$m = e \cdot \theta + \epsilon,$$

where θ is 'ability' and ϵ is a random noise term. An incumbent's ability Θ is a uniform random variable with support $[1, \theta^{max}]$. In periods 0 and 1, there is symmetric incomplete information; neither the individual nor potential employers observe the realisation θ, only the expectation $\mathrm{E}[\theta] = \frac{1+\theta^{max}}{2}$ which we will write as $\overline{\theta}$. By the start of period 2, however, we assume that the individual has learnt her ability and observed her realization θ perfectly. For convenience we assume that ϵ is uniformly distributed on $[\underline{\epsilon}, \overline{\epsilon}]$. All random variables are independent.

In period 0 (pre-intervention), all education employers offer the government teacher wage w^G. In periods 1 and 2 (post-intervention), education wage offers depend on the treatment arm. In FW schools, the wage offer is $w^F > w^G$. In P4P schools, employers offer a performance-contingent contract $\mathcal{C}^P$ that pays $w^G + B$ if $m > \overline{m}$ and w^G otherwise. Employers in the other sector offer a performance-contingent contract in every period. This contract $\mathcal{C}^O$pays w^O if $m > \underline{m}$ and zero otherwise. The compensation schemes available in each period are illustrated in Figure A.1.

In Section A.2, we assume that individuals are free to enter the other sector in any period but can only enter the education sector in period 0 prior to the intervention—the individuals who chose to enter the education sector are the *incumbents* in our empirical study. In Section A.3 we briefly consider the possibility of entry into the education sector in period 1 after the intervention has begun—the individuals who chose to enter the education sector in this period are the *placed recruits* in the empirics.



(a) Period 0

(b) Period 1 and 2

Figure A.1: Compensation schemes

Timing

1. Nature chooses an individual's two-dimensional type. Individuals observe their realization of τ but not their realization of θ.
2. **Period 0 starts.** Education sector employers announce the government wage w^G and number of vacancies n. Employers in the other sector announce a compensation scheme available to all.
3. Individuals choose to apply for a job in the education sector or to seek employment in the other sector.
4. Education sector employers select n applicants to fill the available vacancies. Employers in the other sector hire any individual who applies.
5. Individuals choose effort e_0. A performance metric is realised for each employee. Employers observe these metrics and reward employees in accordance with the compensation scheme announced at Stage 2.
6. **Period 1 starts**. Education sector employers announce a new contract for incumbent teachers, either FW or P4P.
7. Incumbent teachers decide whether to quit for the other sector.
8. Individuals choose effort e_1. A performance metric is realised for each employee. Employers observe these metrics and reward employees in accordance with the compensation scheme announced at Stage 2.
9. **Period 2 starts**. Individuals observe their realization θ. A teacher who experienced P4P in period 1 observes that her intrinsic motivation to teach has shifted down by Δ.
10. Incumbent teachers decide whether to quit for the other sector.
11. Individuals choose effort e_2. A performance metric is realised for each employee. Employers observe these metrics and reward employees in accordance with the compensation scheme announced at Stage 2.

Numerical example At times in the analysis below it will be convenient to use a numerical example. Here we will assume that model parameters take the following values: $\underline{\epsilon} = -5, \overline{\epsilon} = 5, \underline{m} = 1, \overline{m} = 9/2, w^O = 50, w^G = 15, w^F = 30, B = 40$, and $\theta^{max} = 3$.

Appendix A.2 Analysis

Throughout what follows we will assume that individuals make choices myopically, that is, based on current period payoffs only. This simplifies the exposition and, as will become clear below, makes little qualitative difference to the results.[8] This being so, we will break with convention and study period 0 first as it is the simplest case and illustrates key concepts.

Appendix A.2.1 Period 0, pre-intervention

We start at the end of period 0 with effort choices.

[8] Since re-entry into the education sector is not possible, there is an option-value to consider when making occupational choices. However, this merely translates thresholds downwards.

Effort in the education sector An individual recruited to the education sector will choose period 0 effort to solve

$$\max_{e_0} \; w^G - c(e_0, \tau, \text{education}).$$

Even in the absence of extrinsic financial incentives, she exert effort. There is a unique optimal level of effort

$$e_0^G = \tau/2. \tag{8}$$

For effort levels above e_0^G, positive returns are exhausted and effort costs start to kick in.

Effort in the other sector When working in the other sector, an individual chooses period 0 effort to solve

$$\max_{e_0} \; P^O \cdot w^O - c(e_0, \tau, \text{other}),$$

where P^O is the expected probability that m will exceed the threshold $\underline{m}$ given e and $\overline{\theta}$. We can rewrite this probability as

$$\begin{aligned} P^O &= \text{Prob}\left(e \cdot \overline{\theta} + \epsilon > \underline{m}\right) \\ &= \text{Prob}\left(e \cdot \overline{\theta} - \underline{m} > -\epsilon\right) \\ &= \frac{\overline{\epsilon} + e \cdot \overline{\theta} - \underline{m}}{\overline{\epsilon} - \underline{\epsilon}}. \end{aligned}$$

The first order condition for this optimisation problem is

$$\begin{aligned} w^O \cdot \frac{\partial P^O}{\partial e} &= \frac{\partial c}{\partial e} \\ \Rightarrow w^O \cdot \frac{\overline{\theta}}{\overline{\epsilon} - \underline{\epsilon}} &= 2e. \end{aligned}$$

So again there is a unique optimal level of effort

$$e_0^O = \frac{w^O \cdot \overline{\theta}}{2\left(\overline{\epsilon} - \underline{\epsilon}\right)}. \tag{9}$$

Optimal effort under the outside option is (obviously by construction) independent of motivational type τ but instead depends on the outside wage and distributional parameters.

Recruitment into the education sector We now ask which motivational types choose to enter the education sector, having observed the two compensation schemes and anticipating period 0 effort choices. Here, we can define a motivational type τ_0 who, anticipating e_0^G and e_0^O, is indifferent between sectors:

$$\text{E}[u\left(w^G, e_0^G(\tau_0), \tau_0\right)] = \text{E}[u\left(\mathcal{C}^O, e_0^O(\overline{\theta}).\overline{\theta}\right)],$$

It is possible, although not straightforward, to solve for this function τ_0 explicitly. However, things are easier to see in the numerical example, in which case $\tau_0 = 2\sqrt{30}$. Individuals with $\tau \geq \tau_0$ would accept a government wage offer over a job in the other sector. We assume that only these types apply (imagine an infinitesimal application cost) and, moreover, that the government selects n applicants *at random* from this pool. Hence, we can think of $[\tau_0, t^{max}]$ as the type space for incumbent teachers at the start of period 1.

Appendix A.2.2 Period 1

We now undertake our analysis of the first period of the intervention. In modelling terms, all that changes is that, at the start of this period, individuals receive their new contract offers. The contracts are illustrated in Figure A.1: a higher unconditional wage in the FW treatment, and a performance-contingent bonus in the P4P treatment. To ease the exposition, we continue to assume that individuals hold *prior* beliefs over θ.[9]

Effort in the education sector The higher wage *level* in the FW treatment has no bearing on incentives; an incumbent (of a given motivational type) who chooses to accept the FW contract will exert effort

$$e_1^F = \tau/2. \tag{10}$$

In contrast, the performance-contingent scheme announced in the P4P treatment *will* affect incentives. Specifically, an incumbent who chooses to accept the P4P contract (and who does not yet know her ability) will choose effort to solve

$$\max_{e_1} \; P \cdot B + w^G - c(e_1, \tau, \text{education}),$$

where P is the probability that m exceeds the threshold $\overline{m}$ given e and $\overline{\theta}$. Using the same steps as above, it is easy to show that there is a unique optimal level of effort

$$e_1^P = \frac{B \cdot \overline{\theta}}{2\left(\overline{\epsilon} - \underline{\epsilon}\right)} + \frac{\tau}{2}. \tag{11}$$

Effort in the other sector With no new contracts on offer, individuals in the other sector exert the same effort as in period 0: $e_1^O = e_0^O$.

Contract acceptance in the education sector We now ask which motivational types will accept the new contracts announced at the start period 1, again anticipating effort choices later in the period (although not period 2 outcomes). We consider each treatment in turn, starting with FW.

We can define a motivational type τ_1^F who, anticipating e_1^F and e_1^O, is indifferent between sectors:

$$\mathrm{E}[u\left(w^F, e_1^F(\tau_1^F), \tau_1^F\right)] = \mathrm{E}[u\left(\mathcal{C}^O, e_1^O(\overline{\theta}), \overline{\theta}\right)].$$

It is straightforward to show $\tau_1^F < \tau_0$, implying that all (motivational type) incumbents recruited in period 0 who find that their school has been assigned to the FW treatment accept the new contract. This is entirely intuitive—the FW contract offers a strictly higher pay level. Indeed, there are motivational types who would now be willing to enter the education sector, were they able to. Note that, in our numerical example, this threshold is $\tau_1^F = 2\sqrt{15}$.

Analogously, we can define a motivational type τ_1^P who, anticipating e_1^P and e_1^O, is indifferent between sectors:

$$\mathrm{E}[u\left(\mathcal{C}^P, e_1^P(\tau_1^P, \overline{\theta}), \tau_1^P\right)] = \mathrm{E}[u\left(\mathcal{C}^O, e_1^O(\overline{\theta}), \overline{\theta}\right)].$$

In our numerical example, this threshold is $\tau_1^P = 4\sqrt{7} - 8$, and so $\tau_1^P < \tau_1^F < \tau_0$. Hence, (all motivational type) incumbents recruited in period 0 who find that their school has been assigned to the P4P treatment accept the new contract. This is also intuitive, as the P4P constitutes a *bonus* paid on top of the government wage.

[9] A more realistic alternative would be to assume that some (but not full) learning over θ during period 0. Since we are primarily interested in how P4P vs. FW affects retention between periods 1 and 2 (rather than between period 0 and 1), we abstract from learning in period 0.

Predictions Putting all of this together, we see that the intervention does not have a compositional effect in period 1 but it *does* have an incentive effect: the period 1 effort exerted by the average incumbent teacher in a P4P school is higher than the period 1 effort exerted by the average incumbent teacher in a FW school

$$\mathrm{E}[e_1^P] = \frac{B \cdot \overline{\theta}}{2\left(\overline{\epsilon} - \underline{\epsilon}\right)} + \frac{\mathrm{E}[T|T \geq \tau_0]}{2} > \mathrm{E}[e_1^F] = \frac{\mathrm{E}[T|T \geq \tau_0]}{2}.$$

Appendix A.2.3 Period 2

We now turn to analysis of the second period of the intervention. There are two substantive changes. First, individuals have now observed their realization of ability. Second, an incumbent teacher who experienced P4P in period 1 has observed that her intrinsic motivation to teach has shifted down by Δ.

Effort in the education sector In the FW treatment, an incumbent (of a given motivational type) who chooses to remain in the education sector in period 2 will continue to exert the same level of effort as in previous periods

$$e_2^F = \tau/2. \tag{12}$$

In the P4P treatment, an incumbent (of a given motivational type) may now exert more or less effort than in period 1, depending on the news she receives about her ability (and the extent of motivational crowding out). Specifically, in this P4P treatment arm, period 2 effort is given by

$$e_2^P = \frac{B \cdot \theta}{2\left(\overline{\epsilon} - \underline{\epsilon}\right)} + \frac{\tau - \Delta}{2}. \tag{13}$$

Effort in the other sector Individuals in the other sector exert an effort

$$e_2^O = \frac{w^O \cdot \theta}{2\left(\overline{\epsilon} - \underline{\epsilon}\right)}, \tag{14}$$

which also now depends on the realization of ability θ.

Retention in the education sector We complete the analysis by asking which motivational and ability types choose to remain in the education sector (on their post-intervention contract), anticipating subsequent effort choices. Again, we consider each treatment in turn, starting with FW.

For each ability level, we can define a motivational type $\tau_2^F(\theta)$ who, anticipating e_2^F and e_2^O, is indifferent between sectors:

$$\mathrm{E}[u\left(w^F, e_2^F(\tau_2^F), \tau_2^F\right)] = \mathrm{E}[u\left(\mathcal{C}^O, e_2^O(\theta), \theta\right)].$$

In our numerical example, this (ability type-dependent) threshold simplifies to

$$\tau_2^F = \sqrt{5}\sqrt{5\theta^2 - 8},$$

and is illustrated, alongside τ_0 and τ_1^F in Figure A.2 Panel (a). To see the intuition—in particular why $\tau_2^F(\theta)$ is increasing in ability—consider the motivational type τ_0 who is just indifferent between sectors in period 0 when the education wage offer is w^G and she holds the prior belief over ability. In period 1, this type is happy to accept the (more generous) terms of the FW contract. In period 2, if she receives bad news and revises her belief over ability downwards ($\theta < \overline{\theta} = 2$) then a performance-contingent contract is unattractive and she stays put. However, if she receives sufficiently good news ($\theta > 4\sqrt{2/5} \approx 2.5$), then a contract that rewards her ability is attractive and she quits.[10] The same logic applies to more motivated types, except that these types need to receive even better news about their ability before they are willing to forsake their intrinsic motivation to teach and quit for the other sector. For clarity, Figure A.2 Panel (b) shows the outcomes in period 2: the orange-shaded region depicts (τ, θ)-types that are retained in the education sector, while the gray-shaded region depicts (τ, θ)-types that quit.

We now turn to the P4P treatment. For each ability level, we can define a motivational type $\tau_2^P(\theta)$ who, anticipating e_2^P and e_2^O, is indifferent between sectors:

$$\mathrm{E}[u\left(\mathcal{C}^P, e_2^P(\tau_2^P, \theta), \tau_2^P\right)] = \mathrm{E}[u\left(\mathcal{C}^O, e_2^O(\theta), \theta\right)].$$

In our numerical example, this (ability type-dependent) threshold simplifies to

$$\tau_2^P = \sqrt{25\theta^2 + 12} - 4\theta + \Delta,$$

and is illustrated for two values of the crowding-out parameter Δ, alongside τ_0 and τ_1^P, in Figure A.3 Panel (a). The threshold τ_2^P is also increasing in realized ability but at a much slower rate than τ_2^F. This is intuitive: teachers who receive good news about their ability have less reason to quit for the other sector since they are also rewarded for their teaching performance in this P4P treatment arm. So much so, that the two changes that occur in period 2—learning of ability type and motivational crowding out—have no impact on retention. *All* motivational types with a sufficient intrinsic interest in teaching to enter the sector in period 0 on the government wage and willing to stay in period 2.

Predictions Putting the above analysis together, we see that the intervention does have a compositional effect in period 2. For plausible values of Δ, the probability of retaining an incumbent teacher in period 2 is higher under (experienced) P4P than (experienced) FW. Moreover, conditional on retention, the average (expected) ability of an incumbent teacher is higher under (experienced) P4P than (experienced) FW.

In addition to the compositional effect, there is also likely to be an incentive effect. Period 2 effort exerted by the average retained incumbent teacher in a P4P school is

$$\mathrm{E}[e_2^P] = \frac{B \cdot \mathrm{E}[\theta]}{2\left(\bar{\epsilon} - \underline{\epsilon}\right)} + \frac{\mathrm{E}[T|T \geq \tau_0]}{2} - \frac{\Delta}{2}.$$

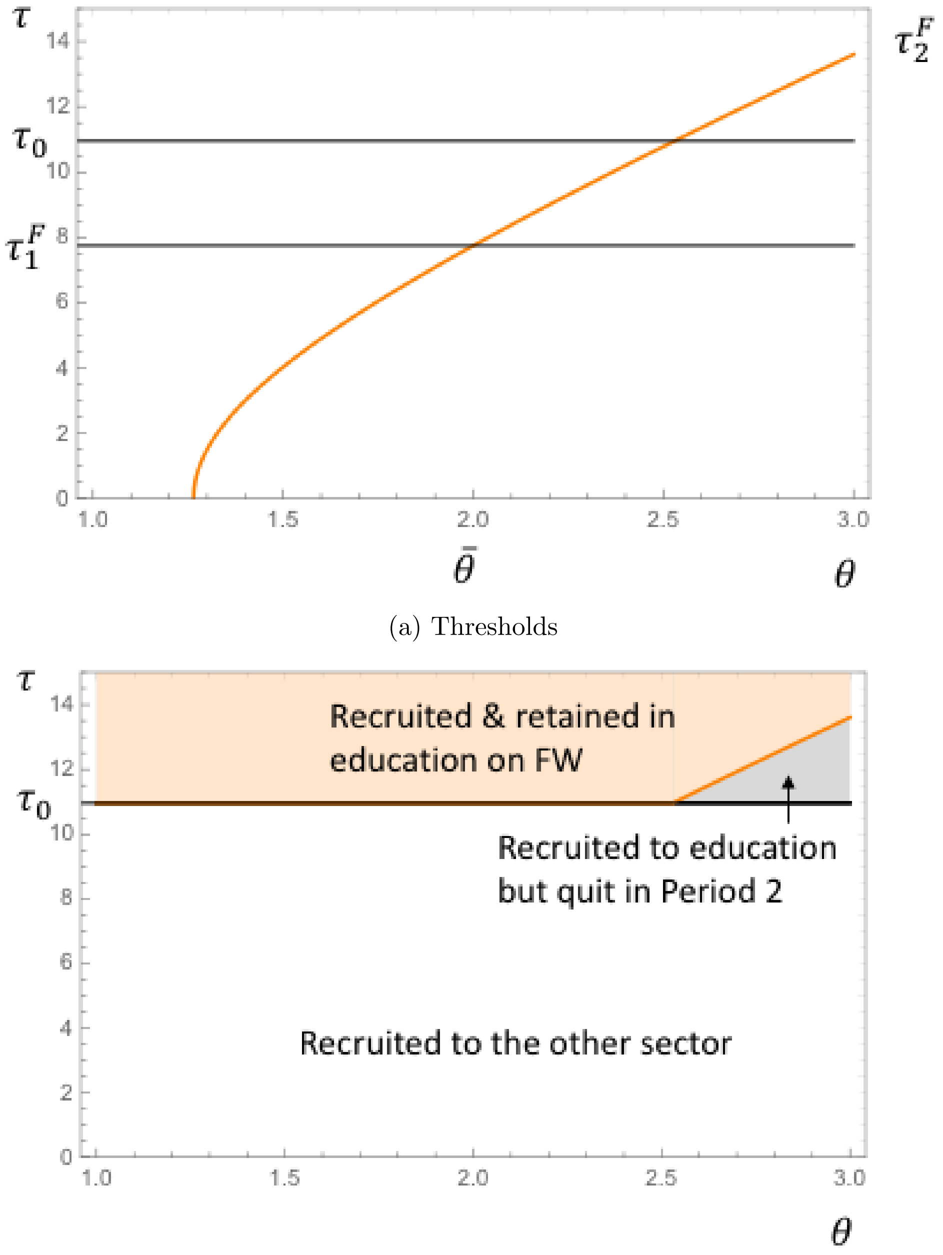
So, on average across P4P schools, effort is lower in period 2 than period 1 by virtue of motivational crowding out.[11] Period 2 effort exerted by the average retained incumbent teacher in a FW school is

$$\mathrm{E}[e_2^F] = \frac{\mathrm{E}[T|T > \tau_2^F(\Theta)]}{2}.$$

The theory does not unambiguously sign this comparison. Effort among retained incumbents will tend to be higher in P4P schools the smaller is (i) the motivational crowding parameter Δ and (ii) the magnitude of the compositional effect (since it is the lower τ types that quit for the other sector).

[10] If $\theta = 4\sqrt{2/5}$, then $\tau_2^F(\theta) = \tau_0$.

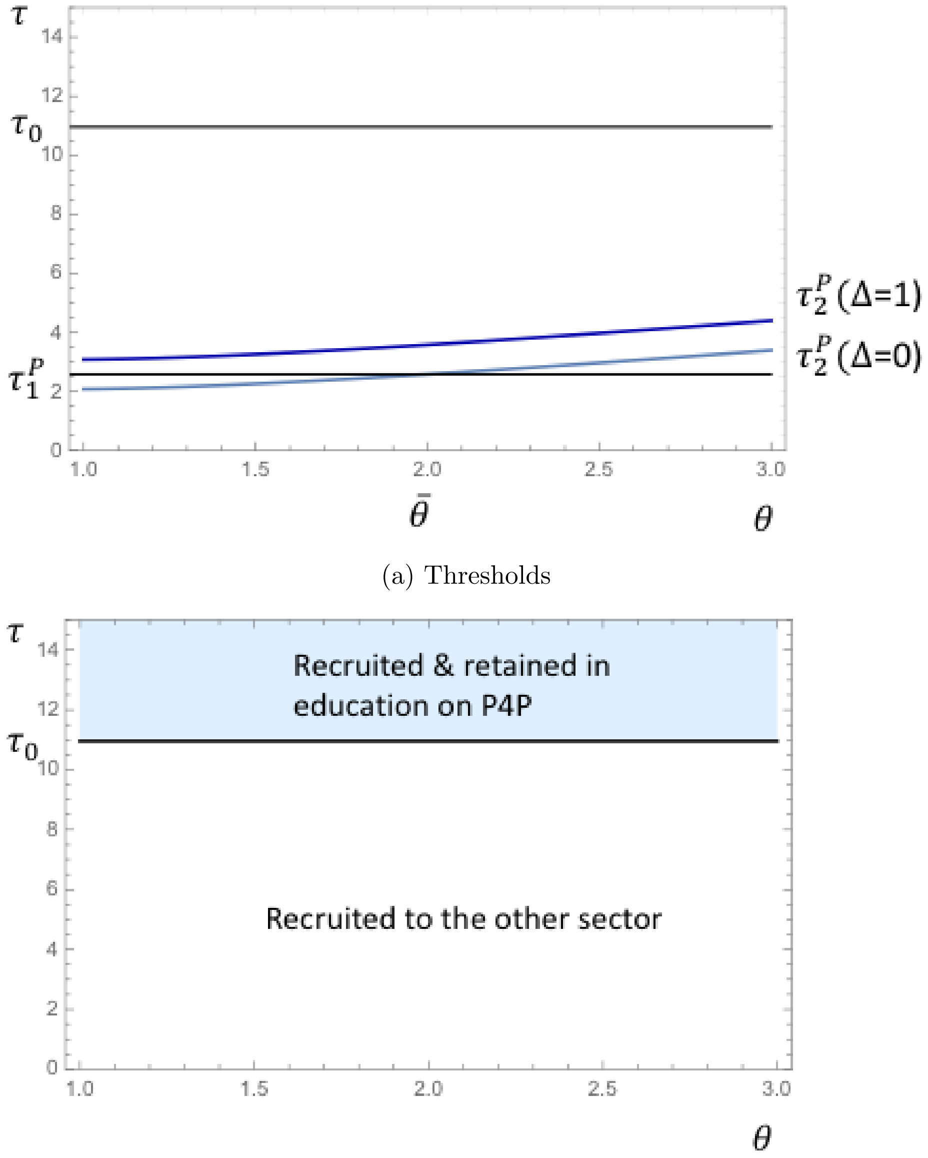
[11] Note that the effort exerted by an individual incumbent may go up or down over time from period 1 to 2. Given the complementarities, good news elicits more effort, and this effect may outweigh the motivational crowding out effect.



(a) Thresholds

(b) Outcomes

Figure A.2: Pre-intervention recruitment & post-intervention retention (of 'incumbent teachers'), FW treatment



(a) Thresholds

(b) Outcomes

Figure A.3: Pre-intervention recruitment & post-intervention retention (of 'incumbent teachers'), P4P treatment

Appendix A.3 Incumbents versus Placed Recruits

The previous section provides predictions for *incumbent* teachers. Power allowing, it is also of interest to compare with predictions for *placed recruits*, i.e. individuals who chose whether to enter the education the sector in period 1 *after* the new contracts have been announced. It is straightforward to see these outcomes in figures above. Motivational types with $\tau \geq \tau_1^F$ are recruited into FW schools in period 1 and are retained in period 2 if $\tau \geq \max\{\tau_1^F, \tau_2^F(\theta)\}$. Hence, the probability of retention is lower for placed recruits than for incumbents. Similarly, motivational types with $\tau \geq \tau_1^P$ are recruited into P4P schools in period 1 and are retained in period 2 if $\tau \geq \max\{\tau_1^P, \tau_2^P(\theta, \Delta)\}$. Again, the probability of retention is lower for placed recruits than for incumbents—now some (τ, θ) types will quit P4P schools for the other sector. The central compositional prediction remains however: for plausible values of Δ the probability of retention is higher under P4P than FW. Moreover, conditional on retention, the average (expected) ability of an incumbent teacher is higher under P4P than FW.