# Monetary Incentives and Image Motivation: Evidence from a Large-Scale Field Experiment with Nigerian Midwives *

Marcus Holmlund†, Pamela Jervis‡, François Maniquet§, Pedro Rosa Dias¶, and Marcos Vera-Hernández‖

July, 2023

**Abstract**

We test the relevance of image motivation in prosocial jobs through a large-scale experiment involving midwives deployed to underserved areas. We randomize incentives to encourage midwife retention and measure midwives' image motivation, and social norms regarding length of service, using lab-in-the-field games. Our motivating behavioral model predicts incentives improve retention, changing the motivational composition of retained midwives by worsening the motivational signal of remaining. Our experimental findings show that incentives improve retention but may backfire as the most highly image-motivated midwives are more likely to leave earlier. Incentives also lengthen the minimum socially acceptable length of service.

*We thank Nava Ashraf, Oriana Bandiera, Erika Deserrano, Ben Elke, Paul Gertler, Matilde Machado, Johan Segers, Jean Tirole for their insightful comments, and seminar participants at IFS, Oxford, UCL, Essex, Surrey, U.C. Louvain, UPF, Madrid, Alicante, Erasmus University, Zurich, IDB, The World Bank, and conference attendees at the 2017 RES conference, the 2020 EEA conference and the 7th Workshop in Behavioral Health Economics. We thank Dr Ugo Okoli, Dr Sidi Ali Mohammed and the SURE-P MCH program team, as well as Arianna Legovini, Muhammad Ali Pate, Hong Wang, Olufemi Adegoke and Emily Crawford for their support. Financial support from the Bill & Melinda Gates Foundation, Strategic Impact Evaluation Fund and The World Bank is gratefully acknowledged. Jervis acknowledges funding from the National Agency for Research and Development (ANID)/Scholarship Program/Doctorado Becas Chile Segunda Convocatoria 2009 - 72101138 and the Center for Research in Inclusive Education (SCIA ANID CIE160009). Maniquet's contribution to this work began during his visit to UCL, funded by the Leverhulme Trust. The experiment is registered in The AEA's RCT Registry with ID AEARCTR-0000617, from where our pre-analysis plan can be downloaded. We obtained IRB approval from the UCL Research Ethics Committee (1827/004) and the National Health Research Ethics Committee of Nigeria (NHREC/01/01/2007).

†World Bank.

‡Universidad de Chile, Institute for Research in Market Imperfections and Public Policy, Institute for Fiscal Studies and Center for Research in Inclusive Education.

§CORE/LIDAM, Université Catholique de Louvain.

¶Imperial College Business School, Imperial College London.

‖University College London and Institute for Fiscal Studies.

## 1 Introduction

Economic incentives lie at the heart of economics. It is well-documented that well-designed economic incentives are generally effective at inducing the exertion of effort as well as hiring productive workers into profit-seeking activities (Lazear (2000), Levitt & Neckermann (2014)). Yet the evidence on their use in pro-social contexts is mixed, with incentives sometimes working, sometimes failing and even backfiring in some situations (Fehr & Falk (2002), Gneezy *et al.* (2011), Kamenica (2012)). The leading explanations for these null or negative effects include incentives influencing the framing of the principal agent-relationship (Gneezy & Rustichini (2000), Heyman & Ariely (2004), Fuster & Meier (2010)), incentives signaling mistrust in the agent (Fehr & Falk (2002), Fehr & List (2004), Falk & Kosfeld (2006), Ellingsen & Johannesson (2007)), incentives signaling unpleasant job attributes (Benabou & Tirole (2003)), and incentives crowding out image motivation (Bénabou & Tirole (2006)).

This paper focuses on the crowding-out of image motivation as proposed by Bénabou & Tirole (2006). They conceptualize motivation as having three components: intrinsic (motivation obtained from the act of giving per se, reflecting a prosocial preference), extrinsic (obtained from any material reward or benefit obtained through giving), and reputational or image motivation (derived from the social approval bestowed by prosocial actions). Bénabou & Tirole (2006) show that, when individuals are heterogeneous in their image motivation, the introduction of incentives may diminish the reputational value of prosocial behavior when incentivized (or even reverse the sign of the signal), as prosocial actions become suspect of being partly or wholly motivated by self-interest. This crowding-out of image motivation is a leading explanation for the mixed evidence on the efficacy of extrinsic incentives in prosocial settings.

We test the relevance of image motivation in a labor market setting by combining lab-in-the-field games with a large-scale experiment in which Nigerian midwives are deployed to work in underserved and often remote rural areas of the country. This setting requires the ongoing exertion of prosocial effort with very significant costs in terms of career progression and living arrangements. Retention incentives are given to a random subsample of midwives in order to encourage them to stay working in the underserved communities for as much time as possible.

In order to motivate, contextualize and explain our empirical findings, we develop a model that includes the main insights from Bénabou & Tirole (2006) but relaxes some of their assumptions and strengthens others in order to fit our setting. Midwives decide how long to stay in the rural areas they were deployed to and their flow of utility has three components. One component is their intrinsic reward, obtained from the act of serving in itself, reflecting a prosocial preference. The importance of this component is heterogeneous among midwives but stable through time. Another component is the material payoff, which is larger if they leave, to the extent that the *opportunity cost* of staying in the village increases with time. This component generalizes the material payoff in Bénabou & Tirole (2006), which they assume to be fixed. The last component is the reputation or image utility, which depends at each moment in time on the pool of midwives who are still in the rural area: the higher the average intrinsic motivation of the pool of midwives in the rural area, the higher will be the image utility of staying in the rural area (as the midwife obtains utility from others inferring that she is highly intrinsically motivated). Through this component, each midwife's decision of either staying or leaving the rural area influences the utility of all other midwives.

The data collected from our large-scale randomized experiment and lab-in-the-field games allow us to test the predictions produced by our model. The model indicates that, under very general assumptions, extrinsic incentives extend, on average, the length of service of midwives in the rural areas they were deployed to. We find support for this prediction, using our collected data on midwives' length of service in the rural areas.

Our model also predicts that there is a threshold of image motivation beyond which extrinsic incentives can cause the most highly image-motivated midwives to be more likely to leave, substantiating the image motivation crowding-out effect. We innovate by directly eliciting image motivation by combining a standard dictator game with a modified one, in which midwives were given a significant sum of money that they were free to split between their personal bank account and a donation to the Red Cross or the Red Crescent. Two versions of the game were played: in one the donation was kept private, while in the other, the donation was made public to all the colleagues of that midwife,in her primary health center. The difference between public and private donations provides a ranking of the importance of image motivation across midwives. Using this novel measure, we find that midwives who decide to leave earlier in the presence of incentives are among those to whom image/reputational motivation is the highest. This empirical result lends support to the image motivation crowding-out hypothesis presented in Bénabou & Tirole (2006).

The model also predicts that extrinsic incentives increase the minimum length of service in rural areas that midwives believe is socially acceptable. We test this prediction by drawing on the incentivized coordination game designed by Krupka & Weber (2013) to elicit midwives' perception of the prevailing norms among their colleagues concerning their minimum socially acceptable length of service in rural areas (for the full sample, at baseline). In consonance with the model predictions, we do find that incentives significantly increase such length of service. Through the lens of our model, we interpret this result as evidence that incentives reduce the average intrinsic motivation in the pool of midwives who stay working in the rural areas, thereby by worsening the motivational signal of remaining[1].

Our paper contributes to several branches of literature. First, it documents the relevance of image motivation crowding-out in a labor market context. Ariely *et al.* (2009) show the relevance of crowding-out of image motivation in the context of charitable donations. Their strategy relies on randomizing whether donations (agents' actions) are public or private. While this strategy is unquestionably neat, it can be challenging to implement it in the context of individuals' day to day job, where it is difficult to keep the agents' actions private from their colleagues[2]. We contribute by proposing a testing strategy that relies on directly measuring image motivation, and randomizing incentives across agents in a labor market setting.

Second, our paper contributes to the important branch of development economics that analyses the role of incentives in the improvement of public service delivery (Finan *et al.* (2017)). This literature can be divided into two strands: one focuses on the impact of incentives on the characteristics of the pool of applicants for a job; the other strand centers on the effect of incentives on the provision of prosocial effort. On the former strand, Deserranno (2019) corroborates Benabou & Tirole (2003) by showing that the announcement of incentives provides a signal to applicants about the characteristics of a job, thereby affecting the pool of applicants. In addition, Ashraf *et al.* (2020) find that, for job applicants with high levels of talent, the offer of career benefits attracts candidates that are more talented, equally prosocial and who perform better at their jobs. The second strand of this literature, concerned with the impact of incentives on the provision of effort, is extensively developed (Lavy (2002), Lavy (2009), Glewwe *et al.* (2010), Muralidharan & Sundararaman (2011), Basinga *et al.* (2011), Duflo *et al.* (2012), Ashraf *et al.* (2014), Behrman *et al.* (2015), Celhay *et al.* (2019), Luo *et al.* (2020), Mohanan *et al.* (2019), among others). Our paper shares similarities with the first of these strands of the literature, although, in our setting, incentives do not affect the pool

[1]This holds in general, except for the improbable scenario where incentives cause the majority of midwives to be more likely to leave.

[2]Implementing the analogue of the experiments run in Ariely *et al.* (2009) in our setting, would have required to publicly announce the names of the midwifes that received a retention incentive. However, in the Nigerian context, and especially in the under-served communities where midwives work, this might have put their safety at risk of theft and violence; we thus decided that it would not be ethical to do this. Moreover, we would not have been able to replicate a private incentive scheme because the midwives from the same Primary Health Center would have learned about it.

of applicants. Instead, they affect the characteristics of the midwives who choose to quit the job earlier, and therefore the composition of the pool of workers at each moment in time.

Our paper also contributes to the literature on social pressure. In Bénabou & Tirole (2006), individuals concerned with their image and social approval by a given social group are exposed to social pressure from that group. Recent contributions corroborate the importance of image motivation and social pressure as drivers of prosocial behavior in a variety of settings: charitable giving (DellaVigna *et al.* (2012)), voting and political participation (Funk (2010), DellaVigna *et al.* (2016), Perez-Truglia & Cruces (2017)), health inputs (Karing (2021a,b), Ashraf *et al.* (2014)) and environmental conservation (Allcott (2011)). Evidence on whether and how social pressure can be shaped by policy is still scant (Bursztyn & Jensen (2017)). By eliciting midwives' perception of the prevailing norms on their minimum acceptable length of service, and showing that this is substantially affected by the provision of extrinsic incentives, we provide new evidence on this knowledge gap.

The remainder of the paper proceeds as follows. Section 2 describes the retention incentive program. Section 3 provides the theoretical model, including its predictions. Section 4 and 5 describe the experimental design and data. Section 6 presents the empirical methods used to test the predictions of the model, and Section 7 the corresponding empirical results. Section 8 provides the conclusions. Appendices provide some of the model technical details and additional tables.

## 2 Deployment of Midwives and Retention Incentives

Nigeria accounts for approximately 2 per cent of the world population, but is burdened with almost 10 per cent of global maternal and neo-natal mortality. One of the leading causes of this has been the severe and systematic lack of midwives in underserved areas of the country. Our experiment takes place in the context of the Subsidy Reinvestment and Empowerment program - Maternal and Child Health Project (SURE-P MCH), a national program designed to redress this situation and implemented by the National Primary Health Care Development Agency (NPHCDA), a parastatal organization under the Nigerian Federal Ministry of Health.

### 2.1 Deployment of Midwives

SURE-P MCH recruited and deployed 1,285 midwives to public Primary Health Centers (PHCs) in underserved areas across all of Nigeria's thirty-six states and the Federal Capital Territory. These facilities were purposely selected based on need and also conditional on meeting minimum infrastructure requirements. The program also included essential upgrading of facilities and the supply of drugs, to facilitate midwives' work in these areas.

Midwife retention was perceived by policy makers as a vital requirement for the success of SURE-P MCH. This can be very challenging given that midwives are deployed to rural areas that are far from where they live and where working conditions are particularly arduous. Thus, in order to promote retention, monetary incentives were given to midwives to encourage them to continue working in the PHCs that they were deployed to, rather than quitting the job and move to an urban center where salaries are generally higher and there are better working/living conditions. Midwives became eligible to receive a retention incentive of NGN 30,000 (roughly USD 200 at the time of the implementation of the program) for each three-months of consecutive attendance at their place of work without unexcused absences. This was paid over and above midwives' regular salaries and any other benefits and corresponded to roughly 25 per cent of the median quarterly salary midwives received from the Federal Government (NGN 40,000 per month). Also, incentives were linked exclusively to attendance, making it feasible to monitor compliance in a setting without a well-functioning clinical information system. Incentives were delivered to qualifying midwives following the completion of each three-month period, and if a midwife failed to comply with attendance at any point during this time, she forfeited the full amount of the incentive. The start and end-date of each three-month period was fixed by SURE-P MCH and so did not vary by midwife. Finally, if a midwife failed to qualify for a round of incentives, she was automatically excluded from all future rounds of incentives.

SURE-P MCH delivered three rounds of monetary incentives corresponding to consistent attendance from December 2013 to February 2014; March to May 2014; and June to August 2014. There was no fixed end-date for the incentive scheme when it was introduced and further rounds of incentives were planned. However, the scheme was cut short due to a budget shortfall related to a sharp fall in international oil prices.

### 2.2 Delivery of the Incentives Contracts

The recruitment of midwives for the program was done without any mention of the retention incentives[3]. These were decided by the Nigerian health authorities and kept confidential from health workers, including PHC managers, until our baseline data collection was completed. As per our pre-analysis plan[4], midwives were

[3] Midwives do not self-select on perceived gains, as in the case of Deserranno (2019).

[4] This can be downloaded from the AEA RCT Registry website: https://www.socialscienceregistry.org/versions/71766/docs/version/document.

randomly assigned to receive either the monetary retention incentives described above, a set of non-monetary retention incentives (special uniforms, calendars and wall clocks), monetary and non-monetary retention incentives simultaneously, or no retention incentives beyond base salary and benefits (pure control).

Two issues arose with non-monetary incentives. First, logistical problems prevented the delivery of non-monetary incentives to midwives. Second, our specific non-monetary incentives were chosen using focus groups discussions in which the midwives proposed different alternatives and ranked them. However, the most preferred alternatives implied some monetary value (training, performance certificates, among others) and hence were discarded. This came at the risk of selecting low powered non-monetary incentives (and indeed our baseline data corroborates that the selected non-monetary incentives were not widely valued by the midwives). Thus, for transparency and consistency with our pre-analysis plan, we report results on non-monetary incentives, but we refrain from providing an interpretation or drawing conclusions about the effectiveness of non-monetary incentives in light of these issues.

Midwives were only informed about their eligibility for incentives at the end of our baseline survey, when they were already working at their PHC, which allows us to abstract from the effect of incentives on selection into the post (as is the case in Deserranno (2019) and Ashraf *et al.* (2020)). At the end of the baseline survey interview, midwives were first shown a video in which the SURE-P MCH Project Director thanked them for their service and announced to them the incentive, or "encouragement" scheme. The Project Director also explained how the scheme was to work and that the incentive was also being offered to SURE-P MCH midwives in other PHCs. A different video was shown depending on the midwife's incentive group. They were then given a contract-letter, customized for each incentive group and signed by the Project Director, explaining the incentive scheme. Midwives assigned to the control group were shown a placebo video thanking them for their crucial services and were also given a corresponding placebo letter.

## 3 Behavioral Model: Extrinsic Incentives, Image Motivation and Social Pressure

In this section we present the model we develop to motivate, contextualize and explain our empirical findings.

### 3.1 Midwives' Utility Function

Let us assume that there is a continuum of SURE-P midwives. Each of them has to decide how long to work in the rural area PHC she was deployed to. When a midwife quits her job at this PHC, she is then free to spend the rest of her professional career in a large city. Formally, each midwife has to choose $v \in [0, 1]$, the share of her career that she spends in the rural area.

Midwives' utility has three components. First, they are interested in their *material well-being*. As long as they stay in the rural area, they get a low flow of material well-being, $y$. When they leave to the city, they get a larger flow of $Y > y$ for the rest of their professional life. These flows are assumed to represent all aspects of their material well-being, including the larger income and the better work and living conditions they would find in the city. We assume midwives are interested in their lifetime material well-being, $c$, which we refer to as income. If they choose to leave the rural area at time $v$, their lifetime income is equal to $c(v) = vy + (1 - v)Y$, from which they derive utility $u(c(v)) = u(vy + (1 - v)Y)$. We assume that $u' > 0$ and $u'' < 0$[5]. If material well-being were the only argument of the preferences, all midwives would spend their entire career in a large city, and a retention scheme, enhancing $y$ to $y'$ would only affect their choice if $y' > Y$, in which case all midwives would stay at their PHC until the end of the retention scheme, which is not what we observe: some midwives stay in the rural area even without the retention scheme.

The second component is the *intrinsic reward* related to their intrinsic motivation, which is the motivation obtained from the act of serving in itself, reflecting a prosocial preference. Working in the rural area, and only in the rural area, gives them a flow of reward proportional to their intrinsic motivation, $\lambda \in [0, L]$, which is assumed to be heterogeneous across the population of midwives. Choosing $v$, therefore, gives a midwife of intrinsic motivation $\lambda$ an intrinsic reward of $v\lambda$.[6] Given that the function $u$, describing how utility is affected by material well-being, is identical across midwives, we should think of $\lambda$ in relative term: a larger $\lambda$ represents a midwife whose intrinsic motivation, relative to her material motivation, is larger. If material and intrinsic reward were the only arguments of the preferences, all midwives would spend some time in the rural area PHC and some time in a large city, with midwives with a larger $\lambda$ staying longer in the rural area. A retention scheme

[5]We show in Appendix A.3 that the assumption of a concave $u$ is consistent with additively separable utility in time.

[6]This way of modeling intrinsic motivation can be viewed as the reduced form of the assumption that working in the rural area gives an intrinsic reward of $\lambda_r$, whereas the reward of working in a city is $\lambda_c < \lambda_r$. The total intrinsic reward is then $v\lambda_r + (1 - v)\lambda_c$, which can be rewritten as $v(\lambda_r - \lambda_c) + \lambda_c$. As the second term is constant, it does not affect choices and can be omitted. By defining $\lambda = \lambda_r - \lambda_c$, we get the current model.

would induce *all* midwives to stay longer in the rural area, which, again, is not what we observe: some midwives leave the rural area earlier under the retention scheme.

The third component is the *reputation/image utility*. Midwives' choices are evaluated by the community of their colleague midwives. We assume the image utility is proportional to the expected intrinsic motivation of those midwives who stay in the rural area. As midwives leave the rural area on a regular basis, this expected intrinsic motivation changes over time. At time $v$, the expected value is $E(\lambda|v; y)$, that is, given the material incentive to stay in the rural area, $y$, and given that some midwives have already left the rural area before time $v$, the community of colleague midwives updates the beliefs about the intrinsic motivation of those who stay. The total image utility for a midwife leaving the rural area after a share $v$ of her career time is as follows:

$$\mu \int_0^v E(\lambda|t; y)dt \tag{1}$$

where $\mu \in [0, M]$ measures how important her image utility is compared to the other components of her utility. Again, $\mu$ should be understood in relationship to the material motivation: a larger $\mu$ represents a midwife whose image motivation, relative to her material motivation, is larger: she dislikes to be viewed as having a low intrinsic motivation or having a large material motivation.

The type of a midwife is, therefore, a pair of intrinsic motivation $\lambda$ and image concern $\mu$. Types are assumed to be distributed according to a distribution function $F$ over $[0, L] \times [0, M]$ with density $0 < f(\lambda, \mu) < \infty$.

To sum up, a midwife of type $(\lambda, \mu)$ chooses to divide her career between staying in the rural area and leaving for the city, $v$, in order to maximize

$$U(v|\lambda, \mu; y) = u(vy + (1 - v)Y) + v\lambda + \mu \int_0^v E(\lambda|t; y)dt. \tag{2}$$

Without loss of generality we normalize the utility by imposing $L = 1$. We further assume $M = 1$ as well, so that the ranges of the marginal affects of both intrinsic and image motivations are the same and equal to $[0, 1]$.

There are three main differences between our model and that of Bénabou & Tirole (2006). First, we only have one image motivation parameter, whereas Benabou and Tirole make the distinction between the utility gain of being thought to be intrinsically motivated and the utility loss of being thought to be extrinsically motivated. As a result, we cannot study policies aiming at influencing the shame of being revealed to be greedy, but nothing in the policy we evaluate is related to such a shame.

Second, and directly related to the first difference, we can dispense with the assumption that midwives' image utility is linear in the time they spend at the village. This is fortunate because in our model in which time plays a crucial role, the marginal image utility depends on the pool of midwives working in the village at a point in time, which has no particular shape and depends on the entire distribution of types.

Third, and more importantly, our reduction to two (instead of three)-dimensional types allows us to dispense with the assumption that intrinsic and image motivations are independently distributed. This implies that some of the predictions we draw from our model (especially Prediction 3 on the backfiring effect of the retention incentives, see below) are more general than the corresponding results in Bénabou & Tirole (2006). This also implies that our last two predictions (Prediction 4 on the efficacy of the retention incentives and Prediction 5 on endogenous social norms, see below) are obtained under an assumption on the distribution of types that could not have been made in the framework of Bénabou & Tirole (2006)'s model and are, therefore, independent of related results in their paper.

### 3.2 Equilibrium

We can think of the decision process as the equilibrium of a game played by this continuum of midwives, who interact with each other only through the image effect (see expression (1)).[7] Let us first characterize the equilibrium. Given that each individual midwife is negligible, she maximizes her utility (2) by taking function $E(\cdot)$ as granted. Given that the domain of $v$ is compact, a global maximum exists as soon as $U$ is continuous. Its only argument that might not be continuous is $E(\cdot)$. This function can only be discontinuous if a positive mass of midwives leave their PHC at the same point in time (that is, if there is bunching at this equilibrium). Predictions 1, 2 and 3 below are valid, should E($\cdot$) be continuous or not.

The first-order condition (FOC) for an interior solution to this problem states that $u'(c(v))c'(v) + \lambda + \mu\partial \int_0^v E(\lambda|t;y)dt/\partial v = 0$, which can be rewritten as

$$u'(vy + (1-v)Y)(Y-y) = \lambda + \mu E(\lambda|v;y). \qquad (3)$$

Corner solutions are obtained when $u'(Y)(Y-y) > \lambda + \mu E(\lambda|0;y)$, in which case the optimal $v$ is equal to 0, or $u'(y)(Y-y) < \lambda + \mu E(\lambda|1;y)$ for the other extreme case.

The left-hand side of the FOC (12) measures the marginal material opportunity cost of not leaving to the city. The right-hand side measures the marginal intrinsic and image rewards of staying in the rural area. At the maximum, the second-order condition (SOC) holds, which reads

$$u''(vy + (1-v)Y)(y-Y)^2 + \mu\frac{\partial E(\lambda|v;y)}{\partial v} \leq 0.$$

[7] A full characterization of the equilibrium can be found in Appendix A.4

The first term of the left-hand side of the SOC is negative, given our assumptions on $u$. The second term, therefore, may be positive or negative, but it cannot be too positive: if the intrinsic motivation of the midwives who stay in the village increases sharply, then midwives don't want to leave at that time.

As a result of these two conditions, if a midwife of type $(\overline{\lambda}, \overline{\mu})$ maximizes her utility function by leaving at time $\overline{v}$, then all midwives of type $(\lambda', \mu')$ such that $\lambda' + \mu' E(\lambda|\overline{v}; y) = \overline{\lambda} + \overline{\mu} E(\lambda|\overline{v}; y)$ also leave at $\overline{v}$, and all midwives of type $(\lambda', \mu')$ such that $\lambda' + \mu' E(\lambda|\overline{v}; y) > \overline{\lambda} + \overline{\mu} E(\lambda|\overline{v}; y)$ stay in the rural area longer than time $\overline{v}$.

This yields the following two empirical predictions:

**Prediction 1:** Let $y$, the incentive scheme, be fixed. For each intrinsic motivation $\lambda$, a midwife with a larger image motivation $\mu$ will stay longer in the rural area.

**Prediction 2:** Let $y$, the incentive scheme, be fixed. For each image motivation $\mu$, a midwife with a larger intrinsic motivation, $\lambda$ will stay longer in the rural area.

### 3.3 Comparative Statics on Material Incentives

We now turn to our main objective, which is to make $y$, the material incentive, vary, and compare equilibria. What is the consequence of an increase in $y$ such as the one caused by the retention incentives given to SURE-P midwives?

We need to analyze the effect of an increase in $y$, say from $y$ to $y'$, on equilibrium condition (12). Let us fix a point in time, $\overline{v}$, and let us examine the change in the pool of midwives who leave at $\overline{v}$ or after. An increase in $y$ unambiguously decreases the left-hand side term of the FOC (12) (both $u'(\overline{v}y+(1-\overline{v})Y)$ and $(Y-y)$ decrease). Midwives who precisely leave at $\overline{v}$ are those for which the right-hand side term has decreased by the same magnitude. An immediate consequence is that midwives who are not image-motivated (that is whose $\mu = 0$) find it more interesting to stay longer in the rural area.

We need to distinguish three cases. To identify these cases, let us define $\lambda^M$, the intrinsic motivation of a midwife with the largest image motivation who leaves the rural area at $\overline{v}$ when the incentive scheme is $y$[8], that is:

$$u'(\overline{v}y + (1 - \overline{v})Y)(Y - y) = \lambda^M + ME(\lambda|\overline{v}; y).$$

Note that $\lambda^M$ is, therefore, the lowest intrinsic motivation among midwives who leave at $\overline{v}$ or after when the incentive scheme is $y$.

Case 1: $E(\lambda|\overline{v}; y') \geq E(\lambda|\overline{v}; y)$. This case is illustrated for $\overline{v} = v^1$ in Figure 1. The change in the expected $\lambda$ is represented by the change in the slope of the

[8] Recall that $M$ stands for the largest possible value of $\mu$.

separating line between types who leave before (in the South-West) and after (in the North-East) $v^1$, as the slope of this line is $\frac{-1}{E(\lambda|\cdot;\cdot)}$. In this case, the combination of the two effects, midwives with $\mu = 0$ stay longer and $E(\lambda|\overline{v}; y)$ increases, imply the number of midwives who leave at $\overline{v}$ or after unambiguously increases (because the set of their types is larger, with respect to inclusion, than the corresponding set of types before the increase in $y$).

Case 2: $E(\lambda|\overline{v}; y') < E(\lambda|\overline{v}; y)$ and $\lambda^M + ME(\lambda|\overline{v}; y') \geq u'(\overline{v}y' + (1-\overline{v})Y)(Y - y')$. This case is illustrated for $\overline{v} = v^2$ in Figure 1. In this case, the pool of midwives still active in the rural area at $\overline{v}$ has a lower average intrinsic motivation than before, with the outcome that the image reward is lower, but this decrease is more than compensated by the increase in material utility, even for the least intrinsic motivated midwives among those who stay. Again, the number of midwives who leave at $\overline{v}$ or after unambiguously increases.

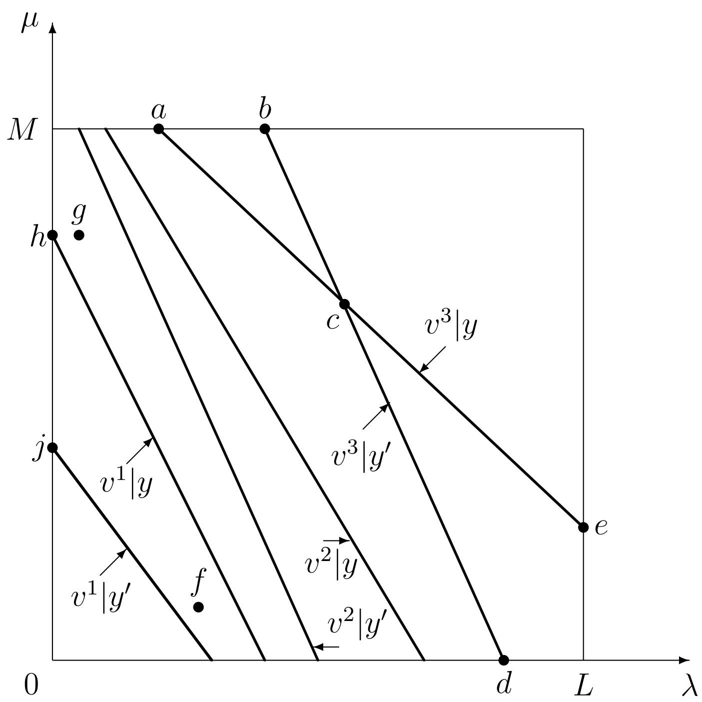



Figure 1: Effect of an increase in the material well-being incentive, from $y$ to $y'$, on the equilibrium: the changes in the separating lines at time $v^1$, $v^2$ and $v^3$ illustrate the three possible cases.

Case 3: $E(\lambda|\overline{v}; y') < E(\lambda|\overline{v}; y)$ and $\lambda^M + ME(\lambda|\overline{v}; y') < u'(\overline{v}y' + (1-\overline{v})Y)(Y - y')$. This case is illustrated for $\overline{v} = v^3$ in Figure 1. In this case, the increase in material well-being is not sufficient to convince the least intrinsically motivated midwives among those who leave at $\overline{v}$ to stay that long after the increase in $y$. This is when the increase in the incentive scheme backfires: some midwives decide to leave earlier. Note that these midwives are also those with the largest image motivation.

We translate Case 3 into the following prediction.

**Prediction 3:** An increase in material incentives may backfire for highly image-motivated midwives: for each intrinsic motivation $\lambda$, there is a threshold in image motivation $\mu(\lambda) \leq M$ such that all midwives with $\mu > \mu(\lambda)$ prefer to leave the rural area earlier under a higher material incentive.

We should add two important remarks regarding Prediction 3. First, the threshold may itself be equal to $M$, in which case there is no backfiring for midwives with this $\lambda$. In this case, all midwives stay longer in the rural area under the retention scheme, and the most image motivated midwives stay longer because it takes them more time to signal that they are intrinsically motivated. Second, we know from the equilibrium characterization that, for any fixed intrinsic motivation $\lambda$, midwives with larger image motivation stay longer in the rural area. This suggests that we could also observe that backfiring impacts more those midwives who stay longer.Third, the fact that the most image motivated midwives leave their PHC earlier under the retention scheme does not mean that they leave it before less intrinsically motivated midwives with the same image concern. On the contrary, Predictions 1 and 2 continue to hold at the new equilibrium, but the additional time spent in the rural area by more image motivated midwives is now shorter.

### 3.4 Overall Effect of Material Incentives

Given the possibility of a backfiring effect for highly image-motivated midwives (Prediction 3), it is crucial to understand what will happen to the number of midwives still active in the rural area at $\overline{v}$ after the increase in $y$. If this number increases, then incentives work: more midwives stay in the rural area. If this number decreases, then the backfiring effect dominates, and material incentives can be counterproductive.

As a function of the joint distribution of types, both cases are possible. The former case, however, is more likely than the latter one. Indeed, if we add the following assumption on the distribution of types, we can prove that incentives work. This assumption captures the idea that the role played by intrinsic motivation is sufficiently important. More precisely, staying in the rural area is determined by a combination of intrinsic and image motivations. We can think of $a\lambda + (1-a)\mu$ as an index of the total motivation of a midwife of type $(\lambda, \mu)$. By restricting our attention to midwives for whom $a\lambda + (1-a)\mu \geq k$, we select midwives that are sufficiently motivated. If we further restrict our attention to $a\lambda + (1-a)\mu \geq k'$, with $k' > k$, we select midwives that are even more motivated. Our assumption requires that the expected intrinsic motivation is also larger in this second group. In other words, we restrict our attention to a population of midwives in which total and intrinsic motivation are positively correlated.

The strength of this assumption in our model depends on the value of $a$. In the case $a = 1$, the restriction is vacuously satisfied by all distributions, because it amounts to looking at expected $\lambda$ when $\lambda > k'$ rather than when $\lambda > k$. If $a$ tends to 0, then this restriction implies that $\lambda$ is positively correlated to $\mu$ which is a restrictive assumption we don't want to make. Therefore, we require that the assumption holds for a range of $a \in [\frac{1}{2}, 1]$.

**Assumption A**: Total and intrinsic motivations are positively correlated. For any value of $a \in [\frac{1}{2}, 1]$, the expected intrinsic motivation $\lambda$ of midwives whose total motivation $a\lambda + (1-a)\mu$ is larger than a given level $k$ is an increasing function of that level $k$[9].

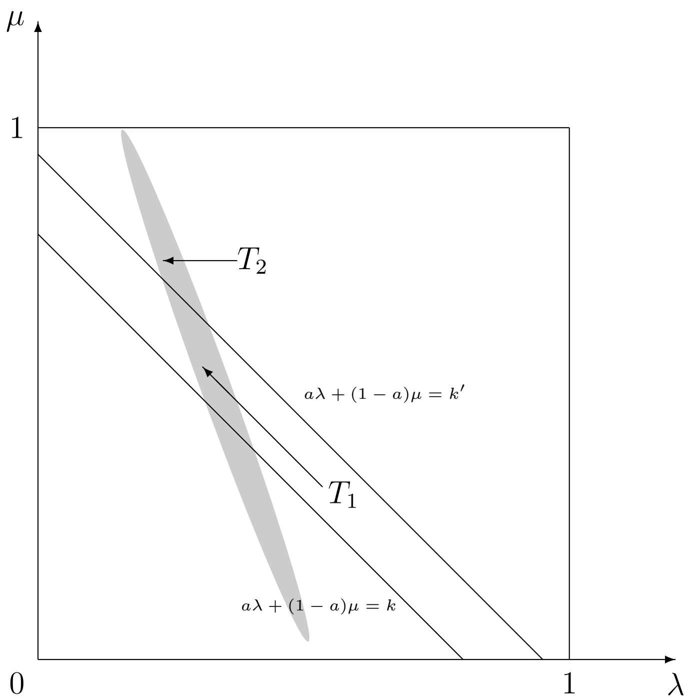



Figure 2: A density function $f(\cdot,\cdot)$ that does not satisfy Assumption A for $a = 0.5$.

Figure 2 illustrates a distribution that does NOT satisfy Assumption A, that is, a distribution under which material incentives may fail to work. Assume almost

[9]Formally, Assumption A is written as follows:

$$\partial \left( \frac{\int_{\max\{0,k-1\}}^{1} \int_{\frac{k-(1-a)\mu}{a}}^{1} \lambda f(\lambda,\mu) d\lambda d\mu}{\int_{\max\{0,k-1\}}^{1} \int_{\frac{k-(1-a)\mu}{a}}^{1} f(\lambda,\mu) d\lambda d\mu} \right) / \partial k \geq 0.$$

all the mass of probability is concentrated in the gray area. Midwives who have a total motivation above $k$ are composed of groups of types $T_1$ and $T_2$. If we now select only those who have motivation above $k'$, then only midwives of types in $T_2$ are left, and they have on average a lower intrinsic motivation than those of types in $T_1$, so that the expected intrinsic motivation has decreased while the expected total motivation has increased. Observe that the problem does not come from $\lambda$ and $\mu$ being negatively correlated. If the gray area was still downward sloping but closer to the horizontal, the problem would disappear (for this value of $a$).

In Appendix 2, we show that when $a = 0.5$, the value of $a$ for which Assumption A is the most restrictive, Assumption A is satisfied for the Farlie-Gumbel-Morgenstern family of density functions (with uniform marginals):

$$f(\lambda, \mu) = 1 - \theta(1 - 2\lambda)(1 - 2\mu)$$

for all $\theta > \underline{\theta} \approx -0.21$ ($\theta$ takes values in $[-1, 1]$). This illustrates that a positive correlation ($\theta > 0$) between $\lambda$ and $\mu$ is sufficient for Assumption A to hold but it is not necessary.

Equipped with Assumption A, we now prove that even in Case 3, the total number of midwives who stay in the rural area increases when material incentives increase. Recall that in this case $E(\lambda|\overline{v}; y') < E(\lambda|\overline{v}; y)$ when $y' > y$. The value of $E(\lambda|\overline{v}; y)$ is the mean between the intrinsic motivation of midwives in two groups: the group of those who leave after $\overline{v}$ with or without the retention incentive, that is those with type above the *bce* line in Figure 1, and the group of those who leave earlier with the retention incentive, that is those with types in the *abc* triangle in Figure 1. Calling 1 (resp. 2) the former (resp. latter) group, we get $E(\lambda|\overline{v}; y) = \frac{p_1}{p_1+p_2}\overline{\lambda}_1 + \frac{p_2}{p_1+p_2}\overline{\lambda}_2$, where $p_1$ and $p_2$ stand for their respective proportion in the total population and $\overline{\lambda}_1$ and $\overline{\lambda}_2$ stand for their average intrinsic motivation. In the same fashion, we compute $E(\lambda|\overline{v}; y') = \frac{p_1}{p_1+p_3}\overline{\lambda}_1 + \frac{p_3}{p_1+p_3}\overline{\lambda}_3$ where group 3 is the set of midwives staying longer with the incentive scheme, that is those with type in the *cde* triangle in Figure 1. The intrinsic motivation of each midwife in group 2 is lower than that of each midwife in group 3, so that $\overline{\lambda}_2 < \overline{\lambda}_3$. Using this relationship, inequality $E(\lambda|\overline{v}; y) > E(\lambda|\overline{v}; y')$ implies $p_3(\overline{\lambda}_1 - \overline{\lambda}_3) > p_2(\overline{\lambda}_1 - \overline{\lambda}_2)$, so it must be the case that $p_3 > p_2$ as long as $\overline{\lambda}_1 > \overline{\lambda}_2$, which is exactly what Assumption A allows us to deduce, as proven in Appendix 2. As a conclusion, those who stay longer than $\overline{v}$ under incentive scheme $y'$ but not under $y$ (group 3) are more numerous than those who stay longer under $y$ than under $y'$ (group 2).

In all three groups, therefore, more midwives are still active in the rural area at time $\overline{v}$ under the retention incentive scheme, for all $\overline{v}$.

Summing the above discussion up, we state the following empirical prediction.

**Prediction 4:** (Under Assumption A:) Incentives work: a larger material incentive, $y$, implies an increase in the average time spent by midwives in the rural area, $v$.

### 3.5 Effect of Material incentives on the Composition of the Pool of Midwives

Figure 1 suggests a further prediction. As illustrated with the case $v^1$, it is possible to have an increase in the average intrinsic motivation of midwives who stay longer than a given amount of time after an increase in the material incentive, but Figure 1 also illustrates that this case is rather unlikely. Indeed, when $E(\lambda|\overline{v}; y') \geq E(\lambda|\overline{v}; y)$, those who are affected the most are low intrinsic motivation midwives (in Figure 1 midwives with $\lambda = 0$ are those for whom the threshold in $\mu$ has decreased the most, from point $h$ to point $j$). Intuitively, adding them to the pool of midwives who stay longer should *decrease* the average intrinsic motivation of that group, contradicting the premise of the argument.

Like for Prediction 4 in Section 3.4, there are joint distributions of $\lambda$ and $\mu$ that allow $E(\lambda|\overline{v}; y)$ to increase with $y$, but these distributions are the exceptions, and, again, Assumption A rules them out, hence our following prediction, the proof of which is provided in Appendix 1:

**Prediction 5:** (Under Assumption A:) At any point in time, the intrinsic motivation of the pool of midwives still active in the rural area is on average lower with the retention scheme: for each $v$, $E(\lambda|\overline{v}; y)$ decreases as $y$ increases.

Prediction 5 is illustrated in Figure 3. Each curve describes the equilibrium relationship between $v$ and $E(\lambda|v; y)$. We know from the analysis of the equilibrium that these curves are increasing. What we have just proven is that the equilibrium curve related to a larger financial incentive $y'$ is everywhere below that related to $y < y'$.

Prediction 5 cannot be directly tested using our experimental data, as we do not observe midwives' intrinsic motivation. What we do observe, however, is the length of stay that midwives find socially acceptable. We will use it to test Prediction 5 indirectly. Indeed, we may assume that each midwife knows that less intrinsically motivated midwives will leave earlier. Therefore, they may consider that a sufficiently long stay reveals a sufficiently high motivation. The length of service that a midwife finds socially acceptable can then be reinterpreted as the time it takes to reveal that they belong to a pool of sufficiently intrinsically motivated midwives. This corresponds to the inverted curves that are displayed in Figure 3: to a given acceptable expected intrinsic motivation $E(\lambda)$ corresponds an acceptable length of stay, which is necessarily larger with the retention scheme.

Hence, we will assess Prediction 5 through testing Prediction 6:

**Prediction 6:** (Under Assumption A:) The minimum length of service in the rural area that midwifes find socially acceptable is on average larger under the retention scheme.

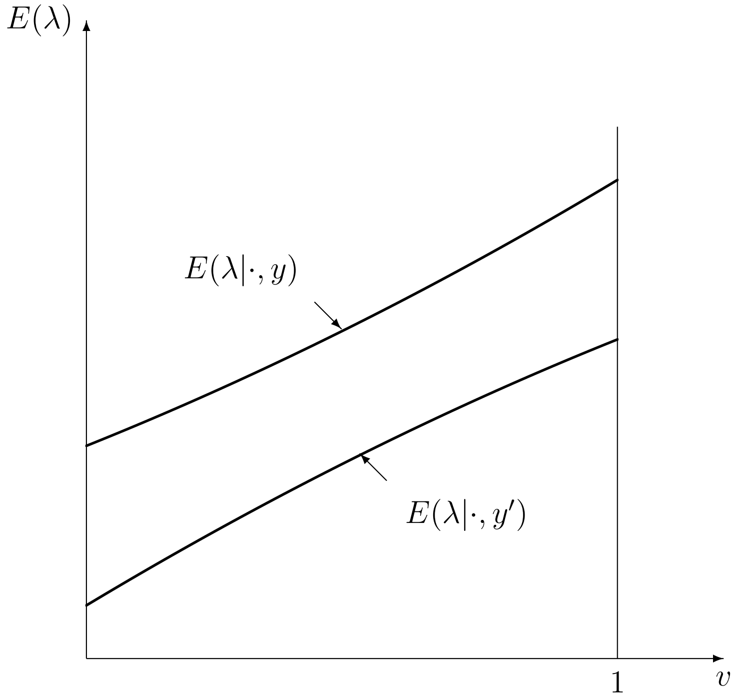



Figure 3: Effect of an increase in the material incentive, from $y$ to $y'$, on the expected intrinsic motivation of midwives who are still in the rural area at $v$.

## 4 Experimental Design

As mentioned above, our experiment takes place in the context of SURE-P MCH, a national program designed to improve maternal and child health in Nigeria. SURE-P MCH grouped PHCs in relative proximity into clusters of four, which were linked to a nearby hospital for secondary care. Our unit of randomization is this cluster of four PHCs. Naturally, the majority of the professional communications between midwives of different PHCs occur within this cluster as they share the same referral hospital. By randomizing at the cluster level rather than at the PHC level, we minimize the possibilities of communication across midwives allocated to different treatment arms.

SURE-P MCH was implemented in 500 PHCs, which corresponds to 125 clusters. We were aware of an independent intervention, the Clinical Governance and Quality Improvement Initiative, designed to improve the management of the PHCs in 24 of the 125 clusters[10] [11]. We built this into our randomization design by grouping these 24 clusters in a large strata: *Strata A*. The 24 clusters were ranked according to the median number of recently hired midwives in each cluster (as reported in administrative records). Randomization blocks of 8 clusters were then formed with clusters whose ranks were 1-8, 9-16, and 17-24. Within each randomization block, two clusters were randomly allocated to each of the treatment arms: Control, Monetary Only incentive, Non-monetary Only incentive, and Both (Monetary and Non-monetary).

Out of the $101 = (125-24)$ remaining clusters, 96 clusters were randomly chosen to comprise *Strata B*, and allocated to 12 blocks of 8 clusters each, following the same process as with *Strata A*. Moreover, as described above, within each randomization block, two clusters were randomly allocated to each of the treatment arms: Control, Monetary Incentive (only), Non-monetary Incentive (only), and Both Incentives (Monetary and Non-monetary). The remaining $5 = (101 - 96)$ clusters, *Strata C*, were randomly allocated to the four treatment arms, with two clusters randomly allocated to the Control arm.

In one of the blocks, there was no baseline data collection in 6 clusters because of violence caused by insurgency that affected the states of Adamawa, Borno, and Yobe, in the north of Nigeria at the time. For this reason, we exclude this entire block from the analysis. Because the randomization was within a block, this does not affect the internal validity of the estimates. Also due to violence, it was not possible to collect baseline data nor deliver the incentive contracts in 5 clusters (in the states of Yobe Borno, Taraba and Zamfara) distributed across 4 blocks. This affected a total of 4 Monetary Incentive (only) arm clusters (19 midwives out of 292), and one cluster from the Non-monetary Incentive (only) arm (4 midwives out of 308), but none from the Control arm nor from clusters with Both Incentives. We treat this small number of midwives as attrition from the experiment, and estimate Lee bounds in Section 7 (Lee (2009)).

## 5 Data

Our data was collected in two rounds: i) baseline, collected between September and November 2013, prior to the start of the midwife retention incentive scheme; ii) endline, collected between December 2014 and February 2015, one year after the introduction of the incentive scheme. At baseline, we interviewed 1,270 midwives.

[10] This intervention is evaluated separately in Dunsch *et al.* (2018).

[11] Our pre-analysis plan mentions that a separate intervention, a Community-based Monitoring Scheme, would be also stratified across the incentives arms. However, the intervention was never rolled out.

The only endline information that we use in the analysis is the date at which the midwives quit their SURE-P job, if they did. We gather this information from all midwives interviewed at baseline, either in-person or by phone, as well as by consulting the administrative records of the PHCs. Table 1 shows that, at baseline, our treatment arms are well balanced in terms of the key covariates despite the attrition of midwives in the Monetary Incentive (only) arm (6.5%) and Non-monetary Incentive (only) arm (1.3%).

Table 1: Balance analysis: Summary of covariates by midwife treatment arms

| | MONETARY | NON-MONETARY | BOTH | NO INCENTIVE | p-value |
|---|---|---|---|---|---|
| Number of Clusters | 27 | 29 | 29 | 27 | |
| Number of midwives | 273 | 304 | 357 | 336 | |
| | mean | mean | mean | mean | |
| ***Covariates of X_ifc (reduced)*** | | | | | |
| Age | 37.2 | 34.3 | 38.9 | 34.3 | 0.134 |
| Years living in community | 3.8 | 3.4 | 4.0 | 3.6 | 0.958 |
| Tenure | 9.7 | 8.4 | 8.0 | 8.9 | 0.566 |
| Rural | 0.6 | 0.7 | 0.6 | 0.5 | 0.357 |
| Midwife experience | 10.8 | 8.4 | 12.4 | 8.1 | 0.210 |
| Maslach Burnout_Emotional Exhaustion | 46.2 | 47.2 | 46.7 | 46.1 | 0.320 |
| Maslach Burnout_Depersonalisation | 28.8 | 29.2 | 29.4 | 29.0 | 0.071 |
| Maslach Burnout_Reduced Personal Accomplishment | 37.6 | 38.3 | 38.1 | 38.3 | 0.270 |
| ***Covariates of X_ifc (extended)*** | | | | | |
| Whether mw's ethnicity is the same as of the catchment area | 0.7 | 0.7 | 0.6 | 0.6 | 0.766 |
| Whether mw's religion is the same as of the catchment area | 0.7 | 0.7 | 0.6 | 0.8 | 0.103 |
| Whether mw's ethnicity is the same as of the area attended primary school | 0.8 | 0.8 | 0.8 | 0.9 | 0.881 |
| Whether mw's religion is the same as of the area attended primary school | 0.9 | 0.9 | 0.9 | 0.9 | 0.840 |
| Whether the mw lived during primary school in a place more rural than the place she lived at baseline | 0.3 | 0.2 | 0.3 | 0.4 | 0.0708 |
| ***Covariates of W_fc (reduced)*** | | | | | |
| Number of midwives | 4.2 | 3.5 | 4.4 | 3.9 | 0.232 |
| Incidence | 0.1 | 0.2 | 0.1 | 0.1 | 0.199 |
| Number of equipment present and working | 15.9 | 16.0 | 15.6 | 16.8 | 0.136 |
| Whether the PHC was included in the Quality Enhancement Program | 0.2 | 0.1 | 0.1 | 0.1 | 0.689 |
| ***Covariates of Z_c (reduced)*** | | | | | |
| Higher grade by midwives | 0.9 | 0.9 | 0.9 | 0.9 | 0.819 |
| Cluster average asset ownership | 0.0 | 0.0 | -0.1 | 0.0 | 0.565 |
| Whether the cluster was part of CCT Program | 0.1 | 0.1 | 0.2 | 0.2 | 0.729 |
| Whether the cluster was part of a stockout intervention | 0.0 | 0.0 | 0.0 | 0.0 | 0.945 |

The baseline collected information through standard survey questions as well as two incentivized behavioral games. These behavioral games are central to our analysis: a novel incentivized game to measure image motivation, which was played before the retention incentive contracts were announced to the midwives; and an incentivized coordination game adapted from Krupka & Weber (2013) to elicit the minimum length of service in the rural PHC that midwives believe is socially acceptable, which was played after midwives became aware of their retention incentives contract.

In order to test Prediction 3, we combine a standard dictator game and a modified one to rank midwives according to the importance they give to image motivation (parameter $\mu$ of utility function (2)). In the standard version of the dictator game, each midwife was asked to divide a fixed sum of money (NGN 2,000 corresponding to 5% of her base monthly salary) between herself and the Nigerian Red Cross/Crescent Society. She was told she could keep all the money for herself, or donate all of it, or divide it between herself and a donation (in increments of NGN 200). The midwife chose one of eleven pre-specified allocations (from keeping the full sum, to donating the full sum), and was reassured that absolutely nobody would know about her choice.

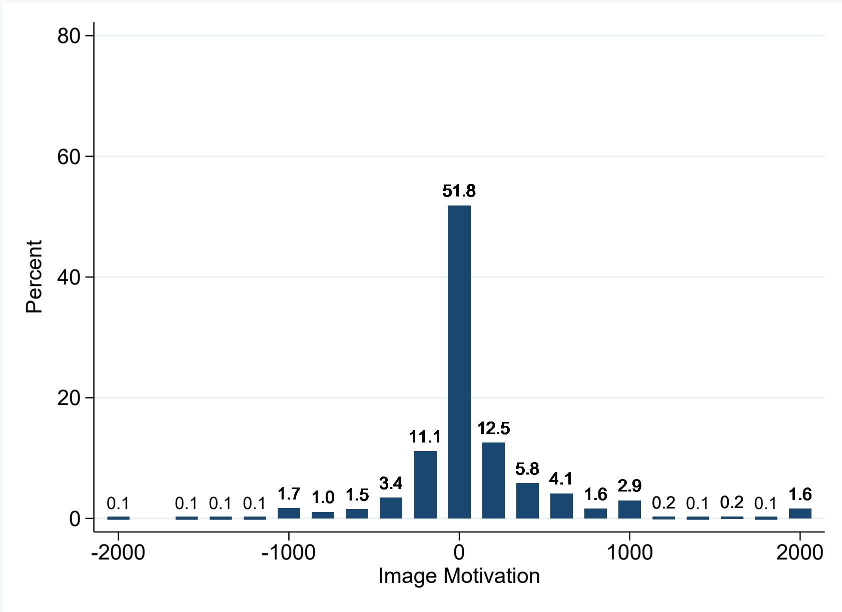
Our innovation was to introduce a "public" version of the dictator game, which was played exactly as the standard (or "private") version but each midwife was told that her choice would be revealed to all her colleague midwives in the PHC. The "public" version was played after the "private" version but the midwife was unaware that there would be a subsequent "public" version when she was playing the "private" one.

Our measure of the importance of image motivation to each midwife, $\mu$, is the difference between the donation in the "public" version of the game and the "private" one. Both donations are affected by the midwife's altruism, but only the "public" one is affected by how much midwives care about their image. By taking the difference between the two, the altruistic element cancels out and the importance of image motivation remains.

Figure 4 shows the distribution of our measure of image motivation. The distribution is Normal-like, with excess mass in zero and more mass on the positive values than the negative ones. Although it is tempting to interpret the negative values of the measure as indicating that the corresponding midwives get disutility from image motivation, it is not straightforward to interpret the value of the social image measure in isolation because the donations are probably affected by the order in which the "public" and "private" versions are played. We kept the same order for all midwives (first the "private" and second the "public") because our objective was not to estimate a measure of image motivation that can be interpreted in isolation, but a measure to rank midwives according to their image motivation.

Table 2 shows the correlation between our measure of image motivation and the other covariates, all at baseline. Image motivation is uncorrelated with age, the number of years the midwife has been living in the community, whether she lived in a rural area before, her experience as midwife and three different measures of burnout. As expected, image motivation is positively correlated with the length of service in the PHC. We also find that midwives who grew up in areas that were more rural than the one where their PHC is located are less image motivated: we

Figure 4: Distribution of Image Motivation




speculate that, for them, embarking on the program and relocating to a rural area can be perceived by their peers as a less costly lifestyle shift than for midwives with a more urban background. Also, we find that those whose religion is the same as the main religion in the catchment area of the PHC are less image motivated: this was expected, as these midwives probably have other opportunities to signal intrinsic motivation (through activities related to their religion, for example).

Finally, it should be mentioned that, corroborating the conclusions from focus group discussions we conducted in 6 Nigerian states prior to the baseline survey, PHC colleagues constitute a closely-knit community and a sound audience for midwives' image concerns: in our baseline data, midwives' median number of friends working with them at the PHC is 6 (92 percent have at least one friend at the PHC), and 35 percent of midwives cohabit with PHC colleagues. Also 94 percent of the midwives' PHC friends are midwives themselves and the median number of midwife friends at their PHC is 4.

In order to test Prediction 6, we have elicited midwives' perception of the minimum length of service in the rural PHC that is socially acceptable. We are interested in knowing the effect of monetary incentives on this perception, so we elicit it after the incentives contract are given to the midwives (or corresponding placebo in the control group). We draw on Krupka & Weber (2013), and ask each midwife whether to quit her job at the SURE-P PHC after a number of months[12] from joining the

[12]The intervals were two months; five months; eight months; eleven months; fourteen months; seventeen months; twenty months; and twenty-four months.

Table 2: Correlation coefficients between image motivation and covariates

| | Coefficient | Standard Error |
|---|---|---|
| Age | 0.056 | 0.045 |
| Years that midwife has been living in the community | -0.065 | 0.042 |
| Length of service in this PHC (in month) | 0.200*** | 0.056 |
| Whether the mw lived in a rural area | 0.040 | 0.056 |
| Number of years working as a midwife (in month) | 0.051 | 0.054 |
| Maslach Burnout Emotional Exhaustion | 0.001 | 0.042 |
| Maslach Burnout Depersonalisation | 0.037 | 0.033 |
| Maslach Burnout Reduced Personal Accomplishment | -0.151 | 0.112 |
| Mw's ethnicity same as catchment area of the PHC | -0.049 | 0.039 |
| Mw's religion same as catchment area of the PHC | -0.128*** | 0.039 |
| Mw's ehtnicity same as area where she attended pre-school | 0.016 | 0.039 |
| Mw's religion same as area where she attended pre-school | -0.057 | 0.047 |
| Mw grew up in more rural place than she is living now | -0.118*** | 0.042 |

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.

program was very socially unacceptable, socially unacceptable, socially acceptable, or very socially acceptable (at baseline, for the full sample). It was made clear to the midwives that their objective was not to express her own preferences but, instead, for her to reveal her beliefs about what her colleagues (in the same treatment arm) thought was socially acceptable. To ensure this, and in line with the approach in Krupka & Weber (2013), midwives were told that, once all surveys had been completed, one of the time periods would be selected at random and that they would receive NGN 1,000 if their response matched the modal response of all SURE-P midwives who had just received the same contract/letter, i.e. if they were able to anticipate the response of the majority of their peers[13].

## 6 Empirical Specification

We estimate the effect of the monetary incentives using the following linear regression estimated using OLS:

$$y_{ifc} = \alpha_0 + \beta_1 MO_c + \beta_2 NMO_c + \beta_3 Both_c + \gamma_1 X_{ifc} + \gamma_2 W_{fc} + \gamma_3 Z_c + \epsilon_{ifc}, \quad (4)$$

where $y_{ifc}$ is an outcome associated to midwife $i$, working at baseline in PHC $f$ from cluster $c$, $MO_c$ takes value 1 if cluster $c$ was randomized into the Monetary

[13] At the time of the survey, NGN 1,000 was the equivalent of half a day's pay. Following the survey, this was transferred to the bank accounts of midwives who correctly chose the modal response.

Only incentive arm and 0 otherwise, $NMO_c$ takes value 1 if cluster $c$ was randomized into the Non-Monetary Only incentive arm and 0 otherwise, and $Both_c$ takes value 1 of cluster $c$ was randomized into the arm that included both monetary and non-monetary incentives. Our specification also includes covariates that vary at the midwife ($X_{ifc}$), health facility ($W_{fc}$), and cluster ($Z_c$) levels. We allow the error term, $\epsilon_{ifc}$ to be arbitrarily correlated amongst midwives working in the same cluster, and hence use standard errors clustered at this level.

A common strategy to improve precision is to include the lagged value of $y_{ifc}$ as covariate in the right hand side of the regression (McKenzie (2012)). However, we cannot follow this strategy because of two reasons: (1) at least for some midwives, the job that they had at baseline was their first job, so they cannot be observed to have left a previous job; (2) some outcomes are measured through experimental games which midwives never played before. Hence, we include a rich set of covariates that vary at the midwife ($X_{ifc}$), health facility ($W_{fc}$), and cluster ($Z_c$) levels to improve the precision of the estimates associated to the experimental arms.

Given the lack of research on midwife attrition, it is unclear what set of covariates will yield substantial gains in precision. In the pre-analysis plan, we specified a reduced (Model A) and a extended (Model B) set of covariates for $(X_{ifc}, W_{fc}, Z_c)$, which are specified at the tables notes[14] [15].

The set of covariates $Z_c$ includes dummy for the strata, and would typically include randomization block dummies. However, given the rich set of covariates we use, we were concerned that this would reduce excessively the available number of degrees of freedom, which are the number of clusters minus the number of covariates. For that reason, instead of using randomization block dummies, we pre-specified a more parametric approach in the pre-analysis plan: to include the interactions between the randomization strata and a quadratic polynomial on the number of recently hired midwives in the cluster, which was the variable used to create the randomization blocks.

In terms of midwife retention, we pre-specified that our main outcome variable, $y_{ifc}$ was whether the midwife was still working in the same health facility as in baseline at 9 months of the incentives contract being delivered to her[16].

[14]The pre-analysis plan was registered before the endline data was available. The pre-analysis plan is downloadable from https://www.socialscienceregistry.org/versions/71766/docs/version/document.

[15]We specified in the pre-analysis plan that if the percentage of missing values for a given covariate was less than 10%, we would replace the missing values with its sample average, but If that percentage exceeded 10%, we would drop the covariate.

[16]We chose nine months because the incentive contracts to Phase 2 midwives was delivered between December 2014 and February 2015, and the follow-up survey for those facilities was planned to take place between October 2015 and December 2015. Hence, we could only guarantee to follow-up Phase 2 midwives for 9 months.

We also estimated a discrete-choice duration model (Jenkins *et al.* (1995)) on the number of months that midwives stayed working in the SURE-P PHC, which avoids specifying a cut-off of minimum duration. We use the same specifications of covariates as in the regression analysis, which we augment with the logarithm of the number of months for the hazard function. This analysis is complemented with non-parametric Kaplan-Meier survival estimates by incentive arm.

## 7 Results

In this section we test empirically the main predictions from our model.

### 7.1 Average Effect of the Retention Incentives

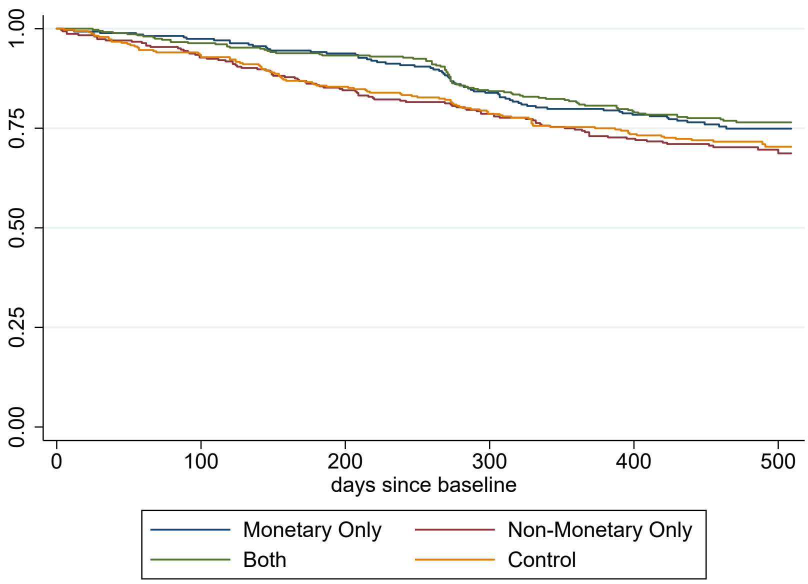
Prediction 4 from our model indicates that monetary incentives increase the average length of service of SURE-P midwives in the PHCs they were assigned to. Figure 5 shows the non-parametric plots (Kaplan-Meier) of the probability of midwives staying in their SURE-P job for the four arms of our trial. This makes three points very clear: (i) the plot of the Monetary Only arm is practically indistinguishable from the plot of the arm with Both incentives (Monetary Only + Non-monetary Only); (ii) likewise, the plot of the Control group is also practically indistinguishable from the plot of the Non-Monetary Only arm; (iii) the probability of midwives staying in their SURE-P job is higher for those in the Monetary Only and Both incentives arms than for those in the Non-Monetary Only and Control arm.

It is not surprising that the Non-Monetary Only arm behaves as the Control arm, and that the Both arm behaves as the Monetary Only arm. As previously indicated, we have reasons to believe that the chosen Non-Monetary incentives were low-powered, and moreover, there were logistical problems preventing the distribution of non-monetary incentives. For these reasons, we do not draw conclusions on the effectiveness of non-monetary retention incentives and only report the results to be loyal to our pre-analysis plan.

Figure 5 also suggests that a narrowing in the gap between the plots takes place approximately 300 days after our baseline data was collected (and the contracts delivered); this coincides with the circulation of (true) information, according to which the SURE-P program would be terminated due to a sharp decrease in the price of oil (which funded the program).

As specified in our pre-analysis plan, we also estimated a parametric discrete-choice duration model (Jenkins *et al.* (1995)), which confirms conclusions gleaned from our Kaplan-Meier estimates; the results of this are shown in Table 3. The probability of leaving the SURE-P job is significantly smaller for midwives in the

Figure 5: Kaplan-Meier Survival Estimates




Monetary Only arm, as well as those in the arm with Both incentives (Monetary + Non-Monetary). It should be noted that our estimated coefficient for the Monetary Only dummy is very similar in size to that of the Both incentives dummy; the same is true for the coefficients of the interaction between the treatment dummies and $\ln(months)$. This goes in line with the overlapping plots reported in Figure (5). The coefficients associated to the Non-Monetary Only arm are much smaller in absolute value and not statistically different from zero, reflecting the overlap with the Control arm that we reported in Figure (5). The positive coefficients of the interaction terms "Monetary Only$*\ln(months)$" and "Both$*\ln(months)$" indicate that the difference between these arms and the Control arm becomes smaller as the number of months after baseline is large (again, as in the Kaplan-Meier plots). Very similar results are obtained when using only two treatment dummies (see Table A.2 in the Appendix).

Further empirical support for Prediction 4 (average positive effect of monetary incentives on midwife retention) is reported in Table 4, which shows the estimates from linear regression (4), where the dependent variable takes value 1 if the midwife quits her job at SURE-P PHC less than nine months after baseline, and 0 otherwise. Again, the results corroborate the main conclusions gleaned from the Kaplan-Meier plots (Figure 5) and the discrete duration model (Table 3). Irrespective of the pre-specified covariates used (Models A and B), Monetary Only incentives decrease by around 6 percentage points the probability that a midwife quits her job in a SURE-P PHC (very similar results are obtained when using only two treatment dummies,

Table 3: Discrete Duration Model Estimates: Probability of Leaving the Job

| | Model A | Model B |
|---|---|---|
| Monetary Only | -1.354** | -1.677** |
| | (0.593) | (0.656) |
| Non-Monetary Only | -0.136 | 0.0879 |
| | (0.358) | (0.421) |
| Both | -1.169*** | -1.440* |
| | (0.447) | (0.761) |
| ln (months) | -0.0353 | 0.0733 |
| | (0.139) | (0.164) |
| Monetary Only * ln(months) | 0.621** | 0.605** |
| | (0.266) | (0.300) |
| Non-Monetary Only * ln(months) | 0.0935 | 0.0218 |
| | (0.177) | (0.199) |
| Both * ln(months) | 0.502** | 0.562 |
| | (0.242) | (0.357) |
| Observations | 18911 | 18895 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.
[a] Model A controls by a reduced version of the midwife characteristics (age, age squared, number of years that midwife has been living in the community and its square term, tenure in the PHC in months and a dummy year, whether the midwife lived in a rural area, number of years working as a midwife (in month) and its square, MBI emotional exhaustion raw score, MBI depersonalisation raw score, and MBI accomplishment raw score), WDC and Facility characteristics (number of midwives working in the PHC at baseline, incidence, number of equipment which are present and working and whether the PHC f was in the quality enhancement program), and socio-economic characteristics (percentage of women who achieved primary education grade 6 or higher per cluster, cluster average asset ownership composite index (a la Anderson 2008), one binary variable for CCT-treatment clusters which were part of an experimental evaluation, another binary variables for cluster in which the CCT was implemented non-experimentally), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.
[b] Model B controls by an extended version of the midwife characteristics (reduced plus whether midwife's ethnicity is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the area where she attended pre-school, and whether the midwife lived during primary school that was more rural than the place where currently living now), and the inclusion of state dummies (state1-state36), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

see Appendix Table A.3). Also in line with previous results, the coefficient associated with the Both (Monetary and Non-Monetary) incentives arm is practically the same as the one with Monetary incentives alone. Overall, our empirical results corroborate Prediction 4 from our model: monetary incentives significantly increased the average length of stay of our SURE-P midwives.

Table 4: Linear model: probability of leaving the job

| | Model A | Model B |
|---|---|---|
| Monetary Only | -0.0561* | -0.0654** |
| | (0.0298) | (0.0250) |
| Non-Monetary Only | 0.0271 | 0.0301 |
| | (0.0302) | (0.0284) |
| Both | -0.0714*** | -0.0695** |
| | (0.0253) | (0.0290) |
| $R^2$ | 0.0725 | 0.116 |
| Observations | 1270 | 1270 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.
[a] Model A controls by a reduced version of the midwife characteristics (age, age squared, number of years that midwife has been living in the community and its square term, tenure in the PHC in months and a dummy year, whether the midwife lived in a rural area, number of years working as a midwife (in month) and its square, MBI emotional exhaustion raw score, MBI depersonalisation raw score, and MBI accomplishment raw score), WDC and Facility characteristics (number of midwives working in the PHC at baseline, incidence, number of equipment which are present and working and whether the PHC f was in the quality enhancement program), and socio-economic characteristics (percentage of women who achieved primary education grade 6 or higher per cluster, cluster average asset ownership composite index (a la Anderson 2008), one binary variable for CCT-treatment clusters which were part of an experimental evaluation, another binary variables for cluster in which the CCT was implemented non-experimentally), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.
[b] Model B controls by an extended version of the midwife characteristics (reduced plus whether midwife's ethnicity is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the area where she attended pre-school, and whether the midwife lived during primary school that was more rural than the place where currently living now), and the inclusion of state dummies (state1-state36), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

As previously indicated, we could not include in the analysis nor deliver the incentive contracts to 6.5% of the midwives from the Monetary Only arm. To assess the robustness of our results, we computed Lee (2009) bounds tightened using the percentage of time that PHCs had functioning electricity[17]. The confidence interval for the lower bound is $(-0.136; -0.034)$ and for the upper bound $(-0.102; -0.006)$. Finally, another piece of evidence that supports our result is the fact that the estimates for Monetary Only and Both are very similar, whilst the Non-Monetary Only seems to mirror the Control arm, due to their low-power and the logistical problems mentioned above.

### 7.2 Crowding-out of Image Motivation

Prediction 3 from our theoretical model indicates that there is a threshold of image motivation beyond which the effect of monetary incentives on retention will be

[17]This information was collected by the SURE-P program when conducting a census of PHCs, and it predates our baseline. The use of the Lee bounds was not pre-registered.

negative. This corresponds to the image motivation crowding-out hypothesis as in Bénabou & Tirole (2006).

To test this prediction from our model, we augment the covariates with an indicator variable which takes value 1 if the midwife's image motivation ($\mu$ in our model), measured by the difference between the public and private donation in our modified dictator game, is larger than a certain threshold; we also include interactions between this image motivation indicator variable and the treatment arm dummies. To increase power (and given that Monetary Only and Both give very similar results in in Figure 5 and Tables 3 and 4), we replace the treatment dummies $MonetaryOnly_c$ and $Both_c$ in regression (4) by a dummy variable ($AnyMonetary_c$) that takes value 1 if cluster $c$ was randomized into the Monetary Only incentive arm or the arm with Both Monetary and Non-monetary incentives, and 0 otherwise. Tables 5 (reduced covariates) and 6 (extended covariates) show the estimates of these models in which the dependent variable is the same as in Table 4, and each column reports the results for a different value of the image motivation threshold (ranging a difference between public and private donation of NGN 0 to NGN 1600)[18].

Tables 5 and 6 make two results clear. First, the top row of the tables indicate that more image-motivated midwives are, on average, less likely to quit the job at a SURE-P PHC. This is consistent with Prediction 1 from our theoretical model and validates our measure of image motivation. Second, and most important, for a sufficiently high threshold of image motivation, i.e. 600 and larger, the coefficient of the interaction term between $AnyMonetary$ and the image motivation dummy is positive and generally statistically significant (from the third to the eight columns in Tables 5 and 6)[19]. Thus, the effect of the monetary incentives on midwife's attrition backfires for highly image motivated midwives, which represent around 11% of the midwives. Overall, these results corroborate Prediction 3, thereby documenting the relevance the image motivation crowding-out hypothesis in our setting.

Figure 6 shows the relevance of image motivation crowding out graphically by plotting the Kaplan-Meier curves by midwives' image motivation. The left panel clearly shows that the $AnyMonetary$ survival curve is above the control one for low image motivated midwives. However, the right panel shows, somehow noisily due

[18]In the pre-analysis plan, we did not anticipate using the regression with the two treatment dummies because we did not necessarily expect the results for the Monetary Only and Both Monetary & Non-monetary to be so similar. The Appendix tables A.4 and A.5 report the results for the pre-specified regression with three treatment dummies, which are similar to the ones reported in Tables 5 and 6 although less precisely estimated.

[19]We report the results for a series of thresholds because the guidance from the theoretical model is that image motivation crowding out takes place when $\mu$ is large enough. At the time of writing the pre-analysis plan we did not have the benefit of the theoretical model and we specified a threshold of zero.

Table 5: Crowding out of Image Motivation (Model A: reduced covariates)

| | IM= $\mathbf{1}[\mu \geq 0]$ | IM= $\mathbf{1}[\mu \geq 400]$ | IM= $\mathbf{1}[\mu \geq 600]$ | IM= $\mathbf{1}[\mu \geq 800]$ | IM= $\mathbf{1}[\mu \geq 1000]$ | IM= $\mathbf{1}[\mu \geq 1200]$ | IM= $\mathbf{1}[\mu \geq 1400]$ | IM= $\mathbf{1}[\mu \geq 1600]$ |
|---|---|---|---|---|---|---|---|---|
| Image Motivation (IM) | -0.0311 | -0.0777 | -0.125** | -0.192*** | -0.176** | -0.373*** | -0.374*** | -0.374*** |
| | (0.0507) | (0.0531) | (0.0485) | (0.0555) | (0.0774) | (0.0511) | (0.0514) | (0.0514) |
| Any Monetary | -0.0569** | -0.0730*** | -0.0782*** | -0.0731*** | -0.0687*** | -0.0679*** | -0.0682*** | -0.0682*** |
| | (0.0276) | (0.0275) | (0.0262) | (0.0250) | (0.0249) | (0.0247) | (0.0247) | (0.0247) |
| Any Monetary*IM | -0.0199 | 0.0560 | 0.130** | 0.144** | 0.114 | 0.342*** | 0.346*** | 0.346*** |
| | (0.0555) | (0.0594) | (0.0602) | (0.0653) | (0.0843) | (0.0716) | (0.0747) | (0.0747) |
| Non-Monetary Only | 0.0147 | 0.0279 | 0.0185 | 0.0186 | 0.0209 | 0.0218 | 0.0241 | 0.0236 |
| | (0.0311) | (0.0320) | (0.0302) | (0.0300) | (0.0300) | (0.0299) | (0.0304) | (0.0304) |
| Non-Monetary Only*IM | 0.0453 | -0.0100 | 0.0834 | 0.145 | 0.156 | 0.426*** | 0.316** | 0.333* |
| | (0.0748) | (0.0695) | (0.0834) | (0.101) | (0.127) | (0.158) | (0.154) | (0.195) |
| $R^2$ | 0.0746 | 0.0759 | 0.0754 | 0.0776 | 0.0755 | 0.0754 | 0.0754 | 0.0753 |
| N | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.
[a] $\mu$ is defined as the difference between public and private donations. The binary variable Image Motivation (IM) denotes whether the difference between public and private donations is greater or equal than a threshold, which varies by column.
[b] Models controls by a reduced version of the midwife characteristics (age, age squared, number of years that midwife has been living in the community and its square term, tenure in the PHC in months and a dummy year, whether the midwife lived in a rural area, number of years working as a midwife (in month) and its square, MBI emotional exhaustion raw score, MBI depersonalisation raw score, and MBI accomplishment raw score), WDC and Facility characteristics (number of midwives working in the PHC at baseline, incidence, number of equipment which are present and working and whether the PHC f was in the quality enhancement program), and socio-economic characteristics (percentage of women who achieved primary education grade 6 or higher per cluster, cluster average asset ownership composite index (a la Anderson 2008), one binary variable for CCT-treatment clusters which were part of an experimental evaluation, another binary variables for cluster in which the CCT was implemented non-experimentally), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

Table 6: Crowding out of Image Motivation (Model B: extended covariates)

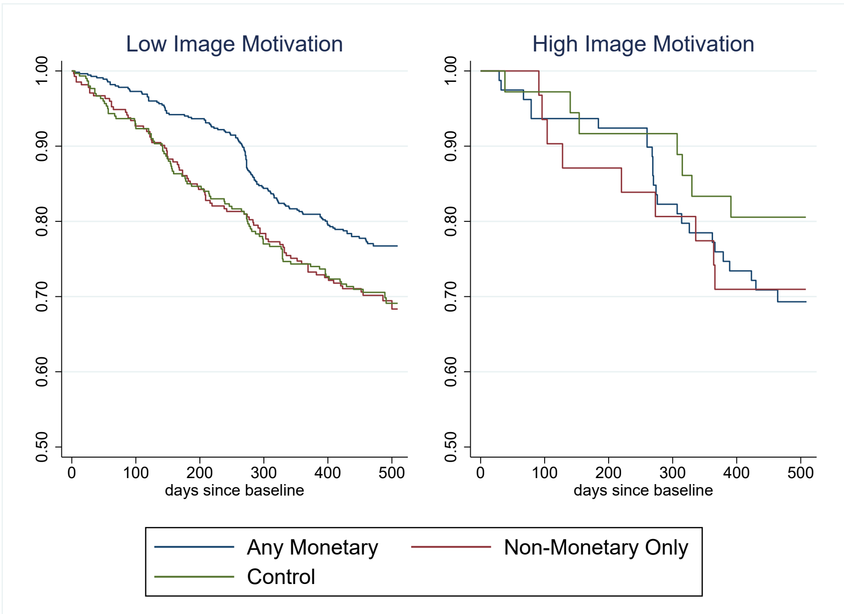
| | IM= $\mathbf{1}[\mu \geq 0]$ | IM= $\mathbf{1}[\mu \geq 400]$ | IM= $\mathbf{1}[\mu \geq 600]$ | IM= $\mathbf{1}[\mu \geq 800]$ | IM= $\mathbf{1}[\mu \geq 1000]$ | IM= $\mathbf{1}[\mu \geq 1200]$ | IM= $\mathbf{1}[\mu \geq 1400]$ | IM= $\mathbf{1}[\mu \geq 1600]$ |
|---|---|---|---|---|---|---|---|---|
| Image Motivation (IM) | -0.0241 | -0.0713 | -0.106** | -0.153*** | -0.136* | -0.324*** | -0.326*** | -0.325*** |
| | (0.0537) | (0.0572) | (0.0442) | (0.0568) | (0.0785) | (0.0528) | (0.0528) | (0.0528) |
| Any Monetary | -0.0581** | -0.0758*** | -0.0792*** | -0.0721*** | -0.0679** | -0.0685*** | -0.0685*** | -0.0685*** |
| | (0.0275) | (0.0273) | (0.0269) | (0.0266) | (0.0265) | (0.0259) | (0.0259) | (0.0259) |
| Any Monetary *IM | -0.0225 | 0.0610 | 0.122** | 0.0976 | 0.0668 | 0.297*** | 0.299*** | 0.298*** |
| | (0.0593) | (0.0632) | (0.0586) | (0.0688) | (0.0866) | (0.0784) | (0.0819) | (0.0820) |
| Non-Monetary Only | 0.0231 | 0.0355 | 0.0246 | 0.0252 | 0.0286 | 0.0269 | 0.0288 | 0.0280 |
| | (0.0310) | (0.0285) | (0.0281) | (0.0284) | (0.0290) | (0.0292) | (0.0291) | (0.0293) |
| Non-Monetary Only*IM | 0.0273 | -0.0344 | 0.0550 | 0.0891 | 0.0919 | 0.353** | 0.255* | 0.291 |
| | (0.0804) | (0.0737) | (0.0821) | (0.0994) | (0.124) | (0.152) | (0.153) | (0.200) |
| $R^2$ | 0.118 | 0.120 | 0.119 | 0.120 | 0.118 | 0.118 | 0.118 | 0.118 |
| N | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.
[a] $\mu$ is defined as the difference between public and private donations. The binary variable Image Motivation (IM) denotes whether the difference between public and private donations is greater or equal than a threshold, which varies by column.
[b] Models controls by an extended version of the midwife characteristics (reduced plus whether midwife's ethnicity is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the area where she attended pre-school, and whether the midwife lived during primary school that was more rural than the place where currently living now), the inclusion of state dummies (state1-state36), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives, and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

to the smaller sample, that the $AnyMonetary$ survival curve is mostly below the control one, consistent with image motivation crowding out.

To allay concerns that the interaction between our measure of image motivation and the treatment arm dummies might be capturing the effect of some other variable correlated with image motivation, we apply LASSO penalized regression to a model which interacts all the individual covariates (the extended version) with the treatment arm dummies. As shown in Figure 7, at all levels of the penalty parameter, the standardized coefficients on the interaction between our image motivation variable ($\mathbf{1}[\mu \geq 600]$) and the $AnyMonetary$ treatment arm is greater than the coefficients of all other interactions. Moreover, increasing the value of the penalty parameter sets the coefficients of all other interaction coefficients to zero before setting the interaction between ($\mathbf{1}[\mu \geq 600]$) and the $AnyMonetary$ treatment arm to zero.

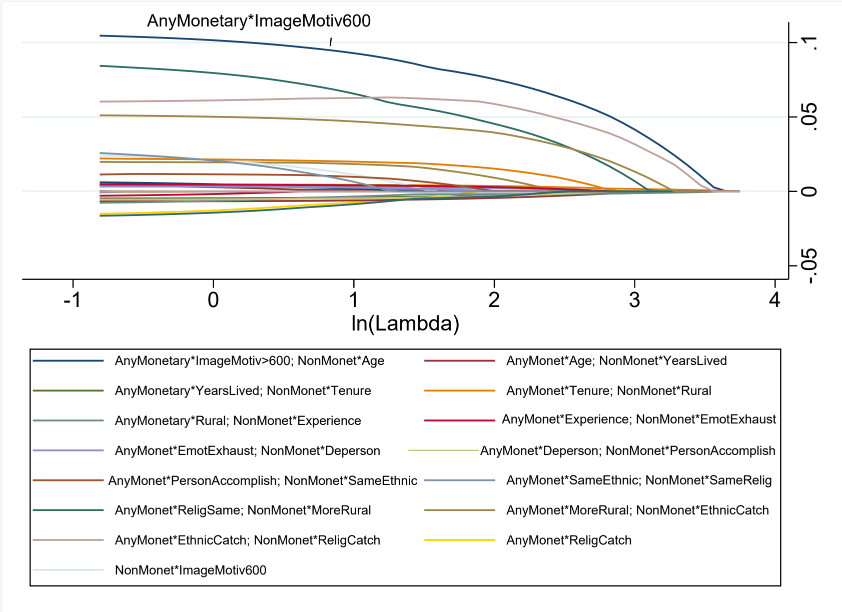
Figure 6: Kaplan-Meier survival estimates by image motivation




## 7.3 Monetary Incentives and the Composition of the Pool of Midwives

In our model, the introduction of incentives simultaneously improves the average length of stay ($v$) but reduces the expected intrinsic motivation ($\lambda$) of the midwives who remain in their PHCs. In other words, when incentives are given, for a midwife's expected $\lambda$ to be the same as before the introduction of such incentives, she needs to stay in her PHC for a longer $v$. In this way, incentives increase midwives' perception of the minimum socially acceptable $v$ (Prediction 6). As explained in Section 3.5, this can also be seen as an indirect test of Prediction 5 according to which, at

Figure 7: Coefficient paths for LASSO regression: Interactions 2-Arms Model




any point in time, the intrinsic motivation of the pool of midwives still working in the SURE-P PHC becomes lower (on average), after midwives are given retention incentives.

In order to empirically test Prediction 6, Table 7 (and Table A.6 in the Appendix) shows the estimated effect of incentives on the minimum socially acceptable length of service in the PHC, elicited at baseline (for the full sample) using the incentivized coordination game we adapted from Krupka & Weber (2013). The dependent variable is the minimum number of months after which midwives believe that most of their fellow midwives consider socially acceptable (or very socially acceptable) to quit the job at the SURE-P PHC. The effect of monetary incentives on this minimum acceptable length of service is positive, large and statistically significant in all our empirical specifications (which vary in the covariates as well as whether we use socially acceptable or very socially acceptable in the definition of the dependent variable). This confirms our Prediction 6 (and therefore also corroborates Prediction 5)[20]. [21]

[20]Although we attempted to test Prediction 5 directly by measuring intrinsic motivation through psychological scales, the responses exhibit little variation and very significant desirability bias, as shown in Table A.1 in Appendix A.5.

[21]Note that, ex-ante, the fact that incentives increase the minimum acceptable length of service, thereby confirming Prediction 6, is not a trivial result. In addition to image concerns, other mechanisms could have been at play here, which could work in opposite directions: for example, by being offered incentives and therefore leaving money on the table when abandoning the program, midwives could have felt more justified to do so. In this context, the positive and significant effect of incentives is in line with the main mechanism at play in our setting being indeed being social image concerns

Table 7: Monetary Incentives and Social Acceptability

| | Model A[a] | Model B[a] | Model A[b] | Model B[b] |
|---|---|---|---|---|
| Any Monetary | 1.520** | 1.083** | 1.309** | 0.943** |
| | (0.649) | (0.508) | (0.508) | (0.443) |
| Non-Monetary Only | 0.866 | 0.824 | 0.595 | 0.419 |
| | (0.833) | (0.557) | (0.598) | (0.398) |
| $R^2$ | 0.0985 | 0.0647 | 0.220 | 0.177 |
| Observations | 1269 | 1269 | 1269 | 1269 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.

[a] The dependent variable is measured as "Earliest month at which it is socially or very socially acceptable to quit the job".

[b] The dependent variable is measured as "Earliest month at which it is very socially acceptable to quit the job".

[c] Model A controls for a reduced version of the midwife characteristics (age, age squared, number of years that midwife has been living in the community and its square term, tenure in the PHC in months and a dummy year, whether the midwife lived in a rural area, number of years working as a midwife (in month) and its square, MBI emotional exhaustion raw score, MBI depersonalisation raw score, and MBI accomplishment raw score), WDC and Facility characteristics (number of midwives working in the PHC at baseline, incidence, number of equipment which are present and working and whether the PHC f was in the quality enhancement program), and socio-economic characteristics (percentage of women who achieved primary education grade 6 or higher per cluster, cluster average asset ownership composite index (a la Anderson 2008), one binary variable for CCT-treatment clusters which were part of an experimental evaluation, another binary variables for cluster in which the CCT was implemented non-experimentally), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

[d] Model B controls for an extended version of the midwife characteristics (reduced plus whether midwife's ethnicity is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the area where she attended pre-school, and whether the midwife lived during primary school that was more rural than the place where currently living now), the inclusion of state dummies (state1-state36), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

## 8 Conclusions

Understanding the mechanisms that cause incentives to work, fail or backfire is a key issue in various areas of economics. Despite the theoretical relevance of the image motivation crowding-out hypothesis (Bénabou & Tirole (2006)), empirical evidence that documents this mechanism directly is scant, especially in prosocial jobs and labor market settings that require the routine exertion of high effort.

We propose a theoretical framework and conduct a large-scale experiment, combined with new lab-in-the-field games, to shed light on these mechanisms in the context of a prosocial job. In our framework, midwives make decisions over their length of service in a setting where these decisions affect the expected intrinsic motivation of the pool of midwives who remain working in their primary health centre. Incentives affect this expectation, thereby impacting the reputational value of the decision to remain working there. This, in turn, influences midwives' choice of length of service in ways that are elucidated in our model.

In our model, the negative impact of monetary incentives on the duration of service in the prosocial job operates through image utility. Our experiment provides evidence consistent with this mechanism: we find that monetary incentives lower the image utility associated with staying, as reflected by a higher minimum socially acceptable length of service (a social norm) perceived by midwives in the presence of monetary incentives.

Our model also predicts that monetary incentives may simultaneously work for some midwives and backfire for others, depending on their type (as captured by preference parameters reflecting image motivation and prosociality). Our experimental results corroborate these prediction. First, they show that monetary incentives succeeded in increasing the average length of service throughout the duration of the program. This is because the main effect of monetary incentives is changing the balance of the decision faced by the majority of midwives between staying in the rural area (being pro-social) or leaving the rural area and enjoying a better material life.

Second, our results also confirm the importance of image motivation, and of the image motivation crowding-out hypothesis. Image-motivated midwives are less likely to leave than other midwives with the same prosocial motivation. Yet the most image-motivated midwives become more likely to quit their prosocial job when they receive monetary retention incentives. These results show that the crowding-out of image motivation by monetary incentives is an important feature of prosocial jobs that require considerable exertion of effort.

Taken as a whole, our results contribute towards a better understanding of the heterogeneity in individual responses to material incentives in prosocial jobs, with implications for policy makers. They show that the crowding out effect is concentrated on highly image motivated individuals.They also show that a simple lab-in-the-field measure of image motivation is predictive of heterogeneous effects of monetary incentives to stay on a prosocial job, in line with Bénabou & Tirole (2006). In addition, our results indicate that incentives can change work-related norms, in our case the minimum socially length of service acceptable by the peers, to the benefit of policy makers. These insights can thus nuance the design of policy in similar contexts to improve overall policy effectiveness.

## References

Allcott, Hunt. 2011. Social norms and energy conservation. *Journal of public economics*, **95**(9-10), 1082–1095.

Ariely, Dan, Bracha, Anat, & Meier, Stephan. 2009. Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American economic review*, **99**(1), 544–55.

Ashraf, Nava, Bandiera, Oriana, & Jack, B Kelsey. 2014. No margin, no mission? a field experiment on incentives for public service delivery. *Journal of public economics*, **120**, 1–17.

Ashraf, Nava, Bandiera, Oriana, Davenport, Edward, & Lee, Scott S. 2020. Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services. *American economic review*, **110**(5), 1355–94.

Basinga, Paulin, Gertler, Paul J, Binagwaho, Agnes, Soucat, Agnes LB, Sturdy, Jennifer, & Vermeersch, Christel MJ. 2011. Effect on maternal and child health services in rwanda of payment to primary health-care providers for performance: an impact evaluation. *The lancet*, **377**(9775), 1421–1428.

Behrman, Jere R, Parker, Susan W, Todd, Petra E, & Wolpin, Kenneth I. 2015. Aligning learning incentives of students and teachers: Results from a social experiment in mexican high schools. *Journal of political economy*, **123**(2), 325–364.

Benabou, Roland, & Tirole, Jean. 2003. Intrinsic and extrinsic motivation. *The review of economic studies*, **70**(3), 489–520.

Bénabou, Roland, & Tirole, Jean. 2006. Incentives and prosocial behavior. *American economic review*, **96**(5), 1652–1678.

Bursztyn, Leonardo, & Jensen, Robert. 2017. Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual review of economics*, **9**, 131–153.

Celhay, Pablo, Martinez, Sebastian, Johannsen, Julia, & Vidal, Cecilia. 2019. Can small incentives have large payoffs? health impacts of a cash transfer program in bolivia.

DellaVigna, Stefano, List, John A, & Malmendier, Ulrike. 2012. Testing for altruism and social pressure in charitable giving. *The quarterly journal of economics*, **127**(1), 1–56.

DellaVigna, Stefano, List, John A, Malmendier, Ulrike, & Rao, Gautam. 2016. Voting to tell others. *The review of economic studies*, **84**(1), 143–181.

Deserranno, Erika. 2019. Financial incentives as signals: experimental evidence from the recruitment of village promoters in uganda. *American economic journal: Applied economics*, **11**(1), 277–317.

Duflo, Esther, Hanna, Rema, & Ryan, Stephen P. 2012. Incentives work: Getting teachers to come to school. *American economic review*, **102**(4), 1241–78.

Dunsch, Felipe, Evans, David K, Macis, Mario, & Wang, Qiao. 2018. Bias in patient satisfaction surveys: a threat to measuring healthcare quality. *Bmj global health*, **3**(2).

Ellingsen, Tore, & Johannesson, Magnus. 2007. Paying respect. *Journal of economic perspectives*, **21**(4), 135–150.

Falk, Armin, & Kosfeld, Michael. 2006. The hidden costs of control. *American economic review*, **96**(5), 1611–1630.

Fehr, Ernst, & Falk, Armin. 2002. Psychological foundations of incentives. *European economic review*, **46**(4-5), 687–724.

Fehr, Ernst, & List, John A. 2004. The hidden costs and returns of incentives—trust and trustworthiness among ceos. *Journal of the european economic association*, **2**(5), 743–771.

Finan, Frederico, Olken, Benjamin A, & Pande, Rohini. 2017. The personnel economics of the developing state. **2**, 467–514.

Funk, Patricia. 2010. Social incentives and voter turnout: evidence from the swiss mail ballot system. *Journal of the european economic association*, **8**(5), 1077–1103.

Fuster, Andreas, & Meier, Stephan. 2010. Another hidden cost of incentives: The detrimental effect on norm enforcement. *Management science*, **56**(1), 57–70.

Glewwe, Paul, Ilias, Nauman, & Kremer, Michael. 2010. Teacher incentives. *American economic journal: Applied economics*, **2**(3), 205–27.

Gneezy, Uri, & Rustichini, Aldo. 2000. Pay enough or don't pay at all. *The quarterly journal of economics*, **115**(3), 791–810.

Gneezy, Uri, Meier, Stephan, & Rey-Biel, Pedro. 2011. When and why incentives (don't) work to modify behavior. *Journal of economic perspectives*, **25**(4), 191–210.

Heyman, James, & Ariely, Dan. 2004. Effort for payment: A tale of two markets. *Psychological science*, **15**(11), 787–793.

Jenkins, Stephen P, *et al.* 1995. Easy estimation methods for discrete-time duration models. *Oxford bulletin of economics and statistics*, **57**(1), 129–138.

Kamenica, Emir. 2012. Behavioral Economics and Psychology of Incentives. *Annual review of economics*, **4**(1), 427–452.

Karing, Anne. 2021a. *The social multiplier from visibility: Experimental evidence from deworming in kenya.* Tech. rept. Mimeo.

Karing, Anne. 2021b. *Social signaling and childhood immunization: A field experiment in sierra leone*. Tech. rept. Mimeo.

Krupka, Erin L, & Weber, Roberto A. 2013. Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the european economic association*, **11**(3), 495–524.

Lavy, Victor. 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of political economy*, **110**(6), 1286–1317.

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

Lazear, Edward P. 2000. The power of incentives. *American economic review*, **90**(2), 410–414.

Lee, David S. 2009. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The review of economic studies*, **76**(3), 1071–1102.

Levitt, Steven D, & Neckermann, Susanne. 2014. What field experiments have and have not taught us about managing workers. *Oxford review of economic policy*, **30**(4), 639–657.

Luo, Renfu, Miller, Grant, Rozelle, Scott, Sylvia, Sean, & Vera-Hernández, Marcos. 2020. Can bureaucrats really be paid like ceos? substitution between incentives and resources among school administrators in china. *Journal of the european economic association*, **18**(1), 165–201.

McKenzie, David. 2012. Beyond baseline and follow-up: The case for more t in experiments. *Journal of development economics*, **99**(2), 210–221.

Mohanan, Manoj, Donato, Katherine, Miller, Grant, Truskinovsky, Yulya, & Vera-Hernández, Marcos. 2019. *Different strokes for different folks: Experimental evidence on the effectiveness of input and output incentive contracts for health care providers with varying skills*. Tech. rept. National Bureau of Economic Research.

Muralidharan, Karthik, & Sundararaman, Venkatesh. 2011. Teacher performance pay: Experimental evidence from india. *Journal of political economy*, **119**(1), 39–77.

Perez-Truglia, Ricardo, & Cruces, Guillermo. 2017. Partisan interactions: Evidence from a field experiment in the united states. *Journal of political economy*, **125**(4), 1208–1243.

# A Appendix

## A.1 Appendix 1. Analytical derivation of Predictions 4 and 5

In this Appendix, we show that Predictions 4 and 5 can be deduced from Assumption A. We reproduce part of Figure 1 below, to illustrate the proof. To complete the derivation of Prediction 4, we need to prove that $\overline{\lambda}_2$, the average intrinsic motivation of midwives whose types lie in the *abc* triangle, is smaller than $\overline{\lambda}_1$, the average intrinsic motivation of midwives whose types lie above the *bce* line.

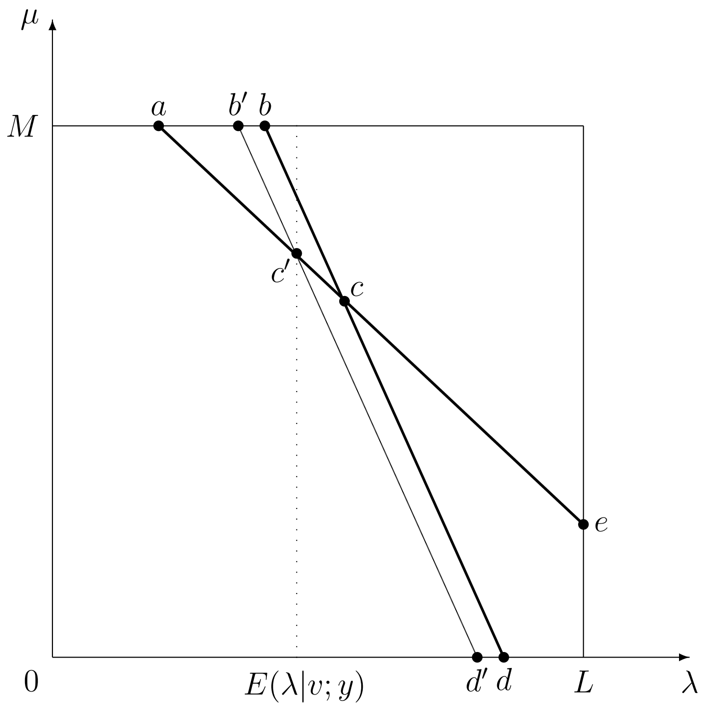



Figure A.8: Derivation of Prediction 4 from Assumption A

Assume, on the contrary, that $\overline{\lambda}_2 \geq \overline{\lambda}_1$. On the graph, the intrinsic motivation of midwives in group 2 are the lowest at point $a$ and the largest at point $c$, so that $\overline{\lambda}_2$ has a value between these two extremes. As a result, given that

$$E(\lambda|v;y) = \frac{p_1\overline{\lambda}_1 + p_2\overline{\lambda}_2}{p_1 + p_2},$$

we have $E(\lambda|v;y) \leq \overline{\lambda}_2$ so that $E(\lambda|v;y)$ also has a value between these two extremes. This is illustrated in Figure A.8, together with a vertical line separating midwives between those who have an intrinsic motivation smaller or larger than $E(\lambda|v;y)$.

Let us recall that the start of the argument is that $E(\lambda|v;y) > E(\lambda|v;y')$. Point $c'$ lies at the intersection between line $b'd'$, parallel to $bd$, and the vertical line at $E(\lambda|v;y)$. Let $\overline{\lambda}'_{13}$ be the average $\lambda$ among the types East of the $b'd'$ line. By Assumption A, $\overline{\lambda}'_{13} < E(\lambda|v;y')$. By transitivity, $\overline{\lambda}'_{13} < E(\lambda|v;y)$, which is impossible, because from the area above the $ae$ line to the area above the $b'd'$ line, we remove types in the $ab'c'$ area, who all have $\lambda \leq E(\lambda|v;y)$ and we add types in the $c'd'e$ area, who all have $\lambda \geq E(\lambda|v;y)$. Consequently, $\overline{\lambda}_2 < \overline{\lambda}_1$, which completes the proof of Prediction 4.

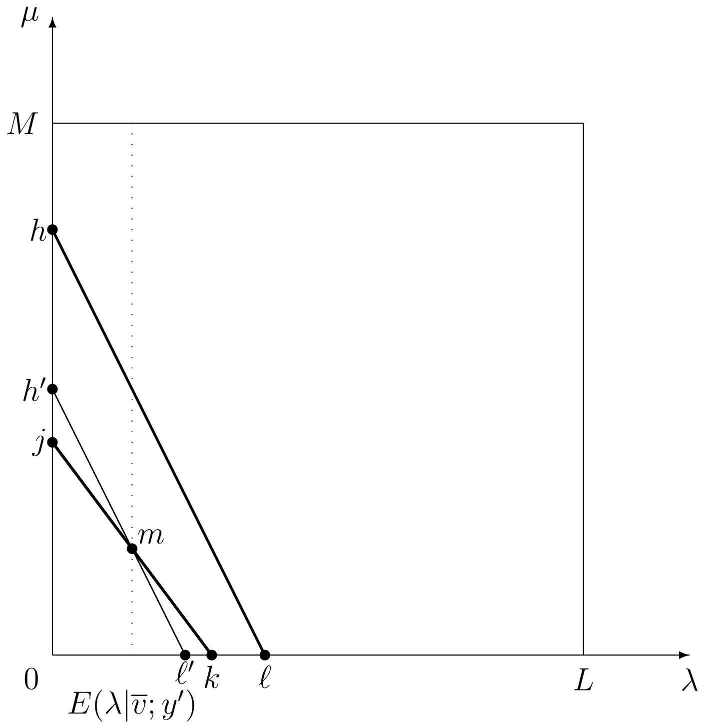



Figure A.9: Derivation of Prediction 5 from Assumption A

To derive Prediction 5, we need to prove that $E(\lambda|\overline{v};y') < E(\lambda|\overline{v};y)$, or, in other words, that case 1 illustrated in Figure 1 is impossible. We reproduce this case in Figure A.9. Let us assume, contrary to what we need to prove, that $E(\lambda|\overline{v};y') \geq E(\lambda|\overline{v};y)$, that is, the average $\lambda$ in the area above the $h\ell$ line is smaller than the average $\lambda$ in the area above the $jk$ line.

Let $\overline{\lambda}_2$ be the average $\lambda$ in the $hjk\ell$ area, that is, $E(\lambda|\overline{v};y')$ is a weighted average between $E(\lambda|\overline{v};y)$ and $\overline{\lambda}_2$. By construction, $\overline{\lambda}_2 \geq E(\lambda|\overline{v};y')$, whereas $\overline{\lambda}_2$ is smaller than the largest value of $\lambda$ in the $hjk\ell$ area, corresponding to point $\ell$. This allows us to position $E(\lambda|\overline{v};y')$ left of $\ell$.

The rest of the proof mimics the derivation of Prediction 4 above. Let $m$ be the point on the $jk$ line that is at the vertical of $E(\lambda|\overline{v};y')$. Let $h'\ell'$ be the parallel to $h\ell$ through $m$. By Assumption A, the average $\lambda$ above the $h'\ell'$ line, let us call it $\overline{\lambda}_1$, is smaller than above the $h\ell$, that is $E(\lambda|\overline{v};y)$. By transitivity, $\overline{\lambda}_1 < E(\lambda|\overline{v};y')$, which is impossible. Indeed, from the area above the $jk$ line to the area above the $h'\ell'$ line, we remove types in the $jmh'$ area, which all have $\lambda < E(\lambda|\overline{v};y')$ and we add types in the $km\ell'$ area, which all have $\lambda > E(\lambda|\overline{v};y')$, so that the average $\lambda$ can only increase.

## A.2 Appendix 2. Assumption A with a Larlie-Gumbel-Morgenstern distribution

In this Appendix, we show that Assumption A is satisfied when $f(\lambda, \mu) = 1 + \theta(1 - 2\lambda)(1 - 2\mu)$ for a wide range of $\theta$ that includes all positive values.

We concentrate on the most restrictive case in the range of parameter $a$, $a = 0.5$. We rewrite the condition on total motivation as $\lambda + \mu \geq k$ (instead of $0.5\lambda + 0.5\mu \geq k$) to save on notation.

We develop the case in which $k \leq 1$. The symmetric case $k \geq 1$ is treated in a similar way.

We need to prove that the derivative of

$$\frac{\int_0^k \int_{k-\mu}^1 \lambda f(\lambda, \mu) d\lambda d\mu + \int_k^1 \int_0^1 \lambda f(\lambda, \mu) d\lambda d\mu}{\int_0^k \int_{k-\mu}^1 f(\lambda, \mu) d\lambda d\mu + \int_k^1 \int_0^1 f(\lambda, \mu) d\lambda d\mu} \tag{5}$$

with respect to $k$ is positive.

We begin with the denominator, $D(k)$. Simple integration computation yields

$$\int \lambda(1 + \theta(1 - 2\lambda)(1 - 2\mu)) d\lambda = (1 + \theta)\frac{\lambda^2}{2} - \theta\mu\lambda^2 - 2\theta\frac{\lambda^3}{3} + 4\theta\mu\frac{\lambda^3}{3}. \tag{6}$$

Taking the value of this integral at the bounds $k - \mu$ and 1, integrating with respect to $\mu$, and inserting the bounds in the resulting equation, we obtain that the first term of $D(k)$ is equal to

$$\frac{(1 + \theta)k}{2} - \frac{2\theta k}{3} + \frac{\theta k^2}{6} - \frac{(1 + \theta)k^3}{6} + \frac{\theta k^4}{4} - \frac{\theta k^5}{15}. \tag{7}$$

In the same way, we compute the second term of $D(k)$, which gives us

$$\frac{1}{2} - \frac{(1 + \theta)k}{2} - \frac{\theta k^2}{6} + \frac{2\theta k}{3} \tag{8}$$

so that, summing up Eqs. 7 and 8,

$$D(k) = \frac{1}{2} - \frac{(1 + \theta)k^3}{6} + \frac{\theta k^4}{4} - \frac{\theta k^5}{15}. \tag{9}$$

The numerator of Eq. 5, $N(k)$, uses first the integral of the density function with respect to $\lambda$,

$$(1 + \theta)\lambda - 2\theta\mu\lambda - \theta\lambda^2 + 2\theta\mu\lambda^2. \tag{10}$$

Taking the bound values, integrating with respect to $\mu$ and taking the values at the bounds for the two terms of the numerator yields

$$N(k) = 1 - \frac{(1 + \theta)k^2}{2} + \frac{2\theta k^3}{3} - \frac{\theta k^4}{6}. \tag{11}$$

The sign of the derivative of Eq. 5 with respect to $k$ is the sign of $D'(k)N(k) -$

$N'(k)D(k)$, that is, using Eqs. 9 and 10,

$$\frac{1+\theta}{2}k-\frac{1+3\theta}{2}k^2+\frac{4\theta}{3}k^3+\frac{(1-\theta)^2}{12}k^4-\frac{\theta+\theta^2}{4}k^5+\frac{13\theta+43\theta^2}{180}k^6-\frac{4\theta^2}{45}k^7+\frac{\theta^2}{90}k^8.$$

Simple computations show that this equation is positive for all $\theta \geq \underline{\theta} \approx -0.21$, for all $k \in [0,1]$.

### A.3 Appendix 3. Model assumptions: Robustness analysis

1. We have assumed that midwives are interested in lifetime income, so that the *opportunity cost* of staying longer in the rural area in terms of material well-being increases. We could have equivalently made the assumption that the job is harder in the rural area, and that how harder it is to work in the rural area increases with age. That is, we could assume that the pain of a unit of labor time in the rural area (resp. the city) is $h$ (resp. $H$), with $h > H$, and there is an increasing and convex disutility function of work so that the material well-being component of global utility becomes:

$$y - d(vh + (1-v)H)$$

in which $y$ is the (identical) income in the rural area or in the city. We would obtain the same comparative static under this set up and the assumption that the incentive to keep midwives in the rural area consists of decreasing $h$.

2. Let us show, by way of a simple example, that the assumption of a concave inter-temporal utility function over consumption is consistent with rational inter-temporal choice. Let us assume first that midwives live longer than their professional career. Let us also assume that they face imperfections on the capital market so that they cannot borrow, when they work in the rural area, and reimburse later when their income has increased. Their life-long material utility can then be written as follows:

$$u(v;y,Y) = \int_0^v \mu(y)dt + \int_v^1 \mu(y_2)dt + \int_1^2 \mu(y_3)dt$$

where $y_2$ and $y_3$ stand for the consumption level when they work at the city and after they stop working respectively, and $\mu$ is the daily utility function of consumption, which is increasing and concave, and where we consider that midwives live one unit of time after their professional career.

The inter-temporal budget constraint writes as follows:

$$(1-v)Y = (1-v)y_2 + y_3.$$

Concavity of $\mu$ implies that $y_2 = y_3$. Let us call it $x$, and we assume that $x > y$, which means that midwives can secure a larger consumption level by migrating to the city, even for the time that follows their professional life. It depends on $v$, according to the following function :

$$x(v) = \frac{1-v}{2-v}Y.$$

We compute the following:

$$\frac{du}{dv} = \mu(y) - \mu(x(v)) - \frac{Y}{2-v}\mu'(x(v)) < 0,$$

where the inequality immediately follows from $\mu(y) < \mu(x(v))$, and

$$\frac{d^2u}{dv^2} = \frac{Y^2}{(2-v)^3}\mu''(x(v)) < 0,$$

corresponding to the two assumptions we imposed on the utility function of total income.

## A.4 Appendix 4. Complete Characterization of the Equilibrium

In this Appendix, we characterize the equilibrium of the game presented in Section 3. The first-order condition reads:

$$u'(vy + (1-v)Y)(Y-y) = \lambda + \mu E(\lambda|v;y). \quad (12)$$

We have to consider two cases, depending on whether $E(\lambda|\cdot;y)$ is a continuous function of $v$.

In case it is continuous, all midwives who leave the village at $v$ are of type $(\lambda, \mu)$ that satisfies Eq. 12, which is linear in $\lambda$ and $\mu$. The second-order condition (SOC), which reads

$$u''(vy + (1-v)Y)(y-Y)^2 + \mu \frac{\partial E(\lambda|v;y)}{\partial v} \leq 0.$$

must also hold, but these two conditions are not sufficient, as the equilibrium may be local and not global. The multiplicity of local maxima determines the nature of the equilibrium when a midwife is indifferent between leaving at two different moments in time, say $v$ and $v'$, $v < v'$, that is when

$$-u'(vy+(1-v)Y)(Y-y)+\lambda+\mu E(\lambda|v;y) = -u'(v'y+(1-v')Y)(Y-y)+\lambda+\mu E(\lambda|v';y).$$

Because $u(\cdot)$ is increasing and concave, $u'(v'y + (1-v')Y) > u'(v'y + (1-v')Y)$, which implies that $E(\lambda|v;y) < E(\lambda|v';y)$: midwives still active in the village are on average more intrinsically motivated at $v'$ than at $v$.

Let $(\lambda', \mu')$ be such that $\lambda' > \lambda$ and $\mu' < \mu$ and

$$\lambda' + \mu' E(\lambda|v;y) = \lambda + \mu E(\lambda|v;y), \quad (13)$$

that is both midwives $(\lambda, \mu)$ and $(\lambda', \mu')$ have a local maximum at $v$. As $\mu' < \mu$, we have $\lambda' + \mu' E(\lambda|v';y) < \lambda + \mu E(\lambda|v';y)$, so that midwife $(\lambda', \mu')$'s marginal utility is negative at $v'$, which implies that she strictly prefers leaving at $v$ than at $v'$: her only global maximum consists of leaving at $v$.

Let $(\lambda', \mu')$ be such that $\lambda' > \lambda$ and $\mu' < \mu$ and

$$\lambda' + \mu' E(\lambda|v';y) = \lambda + \mu E(\lambda|v';y), \quad (14)$$

that is both midwives $(\lambda, \mu)$ and $(\lambda', \mu')$ have a local maximum at $v'$. As $\mu' < \mu$, we have $\lambda' + \mu' E(\lambda|v;y) > \lambda + \mu E(\lambda|v;y)$, so that midwife $(\lambda', \mu')$'s marginal utility is positive at $v$, which implies that she strictly prefers leaving at $v'$ than at $v$: her only global maximum consists of leaving at $v'$.

Let $(\lambda', \mu')$ be such that $\lambda' < \lambda$ and $\mu' > \mu$ and

$$\lambda' + \mu' E(\lambda|v;y) = \lambda + \mu E(\lambda|v;y),$$

that is both midwives $(\lambda, \mu)$ and $(\lambda', \mu')$ have a local maximum at $v$. As $\mu' > \mu$, we have $\lambda' + \mu' E(\lambda|v';y) > \lambda + \mu E(\lambda|v';y)$, so that midwife $(\lambda', \mu')$'s marginal utility is positive at $v'$, which implies that she strictly prefers leaving at $v'$ than at $v$, but she even prefers to stay longer than $v'$: she does not maximize her utility neither at $v$ nor at $v'$. The same can be said of midwives $(\lambda', \mu')$ be such that $\lambda' < \lambda$ and

$\mu' > \mu$ and

$$\lambda' + \mu' E(\lambda|v'; y) = \lambda + \mu E(\lambda|v'; y).$$

To sum up, if midwives of type $(\lambda, \mu)$ are maximizing their utility by leaving either at $v$ or at $v'$, then all midwives $(\lambda', \mu')$ with a larger $\lambda$ for whom Eq. 13 holds leave at $v$, all midwives $(\lambda', \mu')$ with a larger $\lambda$ for whom Eq. 14 holds leave at $v$, and all midwives $(\lambda', \mu')$ with a smaller $\lambda$ leave at a different point in time.

In the case $E(\lambda|\cdot; y)$ is discontinuous at $v''$, it must be the case that a positive mass of midwives leave at $v''$. Let $\overrightarrow{\lambda}$ be the limit of $E(\lambda|\cdot; y)$ for lower value of $v$ and $\overleftarrow{\lambda}$ be the limit of $E(\lambda|\cdot; y)$ for larger value of $v$. It is clear that $\overrightarrow{\lambda} > \overleftarrow{\lambda}$, that is the midwives still active in the village just after $v''$ are on average less intrinsically motivated than those still active in the village just before $v''$. Indeed, if it were not the case, all midwives finding interesting to stay until $v''$ (that is with a positive or zero marginal utility at $v''$ with $E(\lambda|v''; y) = \overrightarrow{\lambda}$) would find it interesting to stay even longer than $v''$ to benefit from this increase in image reward. As a result, no midwife would leave just at $v''$, contradicting the fact that a mass of them leave at $v''$.

A generic equilibrium is represented in Fig. A.10. Each line represents the type of midwives who leave at a particular moment in time. The slope of the line is the opposite of the inverse of the average intrinsic motivation of the midwives still active in the village. Midwives of type $(\overline{\lambda}, \overline{\mu})$ are indifferent between leaving at $v$ and at $v'$, as illustrated by the fact that their type is the intersection between the lines of those who leave at $v$ and at $v'$. A positive mass of midwives leave at $v''$, and the two lines determining the boundary of the types of midwives who leave at $v''$ are represented. Crucially, the upper bound line of these midwives is more vertical than the lower bound: average intrinsic motivation has decreased at $v''$.

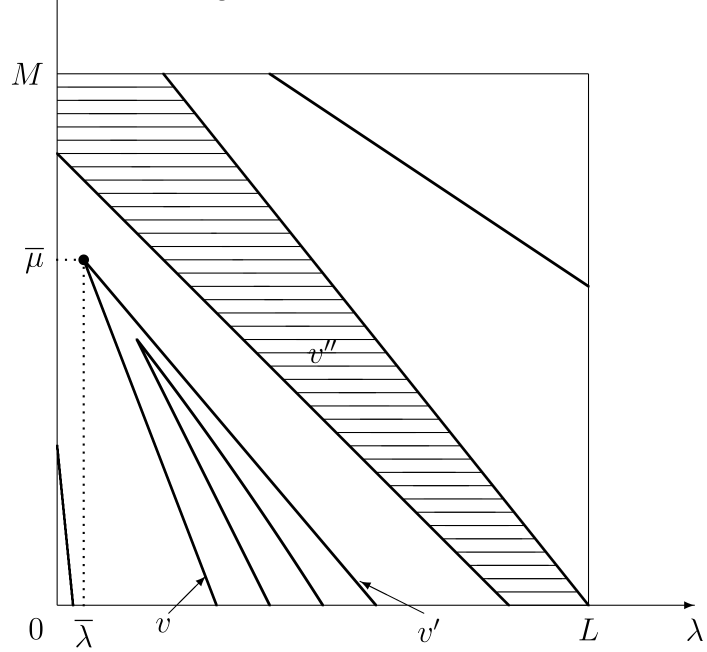



Figure A.10: Characterization of equilibria: midwives $(\overline{\lambda}, \overline{\mu})$ are indifferent between leaving at $v$ and at $v'$; a positive mass of midwives leave at $v''$.

## A.5 Appendix 5. Additional Tables

Table A.1: Prosocial Questionnaire

| | DS | D | NAND | A | AS |
|---|---|---|---|---|---|
| I care about benefiting others through my work | 0.4 | 1.6 | 0.2 | 60.9 | 36.9 |
| I want to have positive impact on others through my work | 0.2 | 0.4 | 0.0 | 53.9 | 45.6 |
| I get energized by working on tasks that have the potential to benefit others | 0.2 | 1.4 | 1.6 | 63.9 | 33.0 |
| I do my best when I'm working on a task that contributes to the well-being of others | 0.2 | 0.2 | 0.4 | 58.8 | 40.5 |
| It is important to me to have the opportunity to use my abilities to benefit others | 0.2 | 0.2 | 0.1 | 59.4 | 40.0 |
| At work, I care about improving the welfare of other people | 0.2 | 0.1 | 0.3 | 60.7 | 38.7 |
| One of my objectives at work is to make a positive difference in other people's lives | 0.2 | 0.2 | 0.1 | 54.8 | 44.8 |
| My work has negative impact on many people | 50.7 | 46.6 | 0.6 | 1.3 | 0.9 |
| I am very aware of the ways in which my work is benefiting others | 0.2 | 1.0 | 6.0 | 73.3 | 19.5 |
| I have negative impact on others in my work on a regular basis | 48.5 | 48.6 | 0.4 | 1.7 | 0.9 |
| I have positive impact on others in my work on a regular basis | 0.4 | 0.3 | 0.8 | 65.9 | 32.6 |
| I feel that other people appreciate my work | 0.2 | 0.2 | 2.7 | 68.0 | 28.9 |
| My work has positive impact on a large number of people | 0.2 | 0.1 | 1.2 | 61.6 | 36.9 |
| I feel that other people respect me for my work | 0.1 | 0.2 | 2.7 | 72.9 | 24.1 |
| I feel that other people value my contributions at work | 0.1 | 0.0 | 1.4 | 71.6 | 26.9 |
| My work really makes others' lives worse | 57.1 | 37.7 | 1.0 | 3.3 | 1.0 |
| The people who benefit from my work are very important to me | 0.6 | 2.3 | 4.1 | 67.2 | 25.7 |
| I care deeply about the people who benefit from my work | 0.3 | 1.1 | 2.6 | 69.7 | 26.3 |

Note: DS: Disagree Strongly / D: Disagree / NAND: Neither Agree Nor Disagree / A: Agree / AS: Agree Strongly

Table A.2: Discrete Duration Model Estimates: Probability of Leaving the Job

| | Model A | Model B |
|---|---|---|
| Any Monetary | -1.245*** | -1.549*** |
| | (0.405) | (0.560) |
| Non-Monetary Only | -0.137 | 0.0893 |
| | (0.359) | (0.419) |
| Time (months) | -0.0347 | 0.0712 |
| | (0.139) | (0.163) |
| Any Monetary * ln(months) | 0.551*** | 0.582** |
| | (0.204) | (0.264) |
| Non-Monetary Only * ln(months) | 0.0931 | 0.0235 |
| | (0.177) | (0.198) |
| Midwives-months | 18911 | 18895 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.
[a] Model A controls by a reduced version of the midwife characteristics (age, age squared, number of years that midwife has been living in the community and its square term, tenure in the PHC in months and a dummy year, whether the midwife lived in a rural area, number of years working as a midwife (in month) and its square, MBI emotional exhaustion raw score, MBI depersonalisation raw score, and MBI accomplishment raw score), WDC and Facility characteristics (number of midwives working in the PHC at baseline, incidence, number of equipment which are present and working and whether the PHC f was in the quality enhancement program), and socio-economic characteristics (percentage of women who achieved primary education grade 6 or higher per cluster, cluster average asset ownership composite index (a la Anderson 2008), one binary variable for CCT-treatment clusters which were part of an experimental evaluation, another binary variables for cluster in which the CCT was implemented non-experimentally), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.
[b] Model B controls by an extended version of the midwife characteristics (reduced plus whether midwife's ethnicity is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the area where she attended pre-school, and whether the midwife lived during primary school that was more rural than the place where currently living now), and the inclusion of state dummies (state1-state36), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.
[c] Both reduced and extended version control by the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

Table A.3: Linear model: probability of leaving the job

| | Model A | Model B |
|---|---|---|
| Monetary (any) | -0.0649*** | -0.0680*** |
| | (0.0247) | (0.0256) |
| Non-Monetary Only | 0.0266 | 0.0298 |
| | (0.0302) | (0.0288) |
| $R^2$ | 0.0723 | 0.116 |
| Observations | 1270 | 1270 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.
[a] Model A controls by a reduced version of the midwife characteristics (age, age squared, number of years that midwife has been living in the community and its square term, tenure in the PHC in months and a dummy year, whether the midwife lived in a rural area, number of years working as a midwife (in month) and its square, MBI emotional exhaustion raw score, MBI depersonalisation raw score, and MBI accomplishment raw score), WDC and Facility characteristics (number of midwives working in the PHC at baseline, incidence, number of equipment which are present and working and whether the PHC f was in the quality enhancement program), and socio-economic characteristics (percentage of women who achieved primary education grade 6 or higher per cluster, cluster average asset ownership composite index (a la Anderson 2008), one binary variable for CCT-treatment clusters which were part of an experimental evaluation, another binary variables for cluster in which the CCT was implemented non-experimentally), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.
[b] Model B controls by an extended version of the midwife characteristics (reduced plus whether midwife's ethnicity is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the area where she attended pre-school, and whether the midwife lived during primary school that was more rural than the place where currently living now), and the inclusion of state dummies (state1-state36), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

Table A.4: Crowding out of Image Motivation (Model A: reduced covariates)

| | IM= $\mathbf{1}[\mu \geq 0]$ | IM= $\mathbf{1}[\mu \geq 400]$ | IM= $\mathbf{1}[\mu \geq 600]$ | IM= $\mathbf{1}[\mu \geq 800]$ | IM= $\mathbf{1}[\mu \geq 1000]$ | IM= $\mathbf{1}[\mu \geq 1200]$ | IM= $\mathbf{1}[\mu \geq 1400]$ | IM= $\mathbf{1}[\mu \geq 1600]$ |
|---|---|---|---|---|---|---|---|---|
| Image Motivation (IM) | -0.0309 | -0.0777 | -0.124** | -0.191*** | -0.175** | -0.371*** | -0.372*** | -0.372*** |
| | (0.0508) | (0.0531) | (0.0485) | (0.0554) | (0.0772) | (0.0506) | (0.0508) | (0.0509) |
| Monetary Only | -0.0383 | -0.0550 | -0.0579* | -0.0608* | -0.0578* | -0.0571* | -0.0574* | -0.0573* |
| | (0.0380) | (0.0342) | (0.0326) | (0.0311) | (0.0309) | (0.0304) | (0.0304) | (0.0304) |
| Monetary Only*IM | -0.0502 | -0.00498 | 0.0183 | 0.0740 | 0.0551 | 0.230*** | 0.230*** | 0.230*** |
| | (0.0657) | (0.0689) | (0.0612) | (0.0666) | (0.0875) | (0.0659) | (0.0662) | (0.0662) |
| Non-Monetary Only | 0.0153 | 0.0287 | 0.0197 | 0.0195 | 0.0216 | 0.0223 | 0.0245 | 0.0240 |
| | (0.0311) | (0.0319) | (0.0302) | (0.0300) | (0.0300) | (0.0298) | (0.0303) | (0.0303) |
| Non-Monetary Only*IM | 0.0458 | -0.00945 | 0.0831 | 0.144 | 0.156 | 0.425*** | 0.314** | 0.332* |
| | (0.0748) | (0.0695) | (0.0833) | (0.101) | (0.126) | (0.158) | (0.154) | (0.194) |
| Both | -0.0706** | -0.0867*** | -0.0934*** | -0.0823*** | -0.0767*** | -0.0759*** | -0.0763*** | -0.0762*** |
| | (0.0270) | (0.0277) | (0.0265) | (0.0255) | (0.0254) | (0.0251) | (0.0251) | (0.0251) |
| Both*IM | 0.00334 | 0.0995 | 0.209*** | 0.191** | 0.153* | 0.427*** | 0.444*** | 0.444*** |
| | (0.0547) | (0.0606) | (0.0672) | (0.0733) | (0.0910) | (0.100) | (0.110) | (0.110) |
| $R^2$ | 0.0754 | 0.0777 | 0.0793 | 0.0787 | 0.0761 | 0.0766 | 0.0767 | 0.0766 |
| Observations | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.
[a] $\mu$ is defined as the difference between public and private donations. The binary variable Image Motivation (IM) denotes whether the difference between public and private donations is greater or equal than a threshold, which varies by column.
[b] Models controls by a reduced version of the midwife characteristics (age, age squared, number of years that midwife has been living in the community and its square term, tenure in the PHC in months and a dummy year, whether the midwife lived in a rural area, number of years working as a midwife (in month) and its square, MBI emotional exhaustion raw score, MBI depersonalisation raw score, and MBI accomplishment raw score), WDC and Facility characteristics (number of midwives working in the PHC at baseline, incidence, number of equipment which are present and working and whether the PHC f was in the quality enhancement program), and socio-economic characteristics (percentage of women who achieved primary education grade 6 or higher per cluster, cluster average asset ownership composite index (a la Anderson 2008), one binary variable for CCT-treatment clusters which were part of an experimental evaluation, another binary variables for cluster in which the CCT was implemented non-experimentally), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

Table A.5: Crowding out of Image Motivation (Model B:extended covariates)

| | IM= $\mathbf{1}[\mu \geq 0]$ | IM= $\mathbf{1}[\mu \geq 400]$ | IM= $\mathbf{1}[\mu \geq 600]$ | IM= $\mathbf{1}[\mu \geq 800]$ | IM= $\mathbf{1}[\mu \geq 1000]$ | IM= $\mathbf{1}[\mu \geq 1200]$ | IM= $\mathbf{1}[\mu \geq 1400]$ | IM= $\mathbf{1}[\mu \geq 1600]$ |
|---|---|---|---|---|---|---|---|---|
| Image Motivation (IM) | -0.0242 | -0.0716 | -0.105** | -0.153*** | -0.135* | -0.323*** | -0.325*** | -0.324*** |
| | (0.0537) | (0.0573) | (0.0443) | (0.0569) | (0.0787) | (0.0522) | (0.0521) | (0.0521) |
| Monetary Only | -0.0482 | -0.0647** | -0.0674** | -0.0686** | -0.0652** | -0.0637** | -0.0642** | -0.0641** |
| | (0.0305) | (0.0273) | (0.0272) | (0.0264) | (0.0263) | (0.0254) | (0.0255) | (0.0255) |
| Monetary Only *IM | -0.0477 | 0.00828 | 0.0225 | 0.0506 | 0.0186 | 0.187** | 0.188** | 0.188** |
| | (0.0700) | (0.0726) | (0.0631) | (0.0789) | (0.0994) | (0.0799) | (0.0797) | (0.0797) |
| Non-Monetary Only | 0.0235 | 0.0364 | 0.0266 | 0.0257 | 0.0287 | 0.0273 | 0.0290 | 0.0282 |
| | (0.0308) | (0.0282) | (0.0280) | (0.0281) | (0.0287) | (0.0289) | (0.0288) | (0.0290) |
| Non-Monetary Only*IM | 0.0276 | -0.0338 | 0.0548 | 0.0892 | 0.0915 | 0.352** | 0.254* | 0.290 |
| | (0.0804) | (0.0738) | (0.0822) | (0.0996) | (0.124) | (0.152) | (0.153) | (0.200) |
| Both | -0.0651** | -0.0833*** | -0.0882*** | -0.0751** | -0.0705** | -0.0727** | -0.0729** | -0.0729** |
| | (0.0307) | (0.0310) | (0.0304) | (0.0302) | (0.0299) | (0.0294) | (0.0293) | (0.0293) |
| Both*IM | -0.00272 | 0.0983 | 0.192*** | 0.129* | 0.0991 | 0.383*** | 0.395*** | 0.394*** |
| | (0.0582) | (0.0654) | (0.0655) | (0.0721) | (0.0887) | (0.0932) | (0.101) | (0.101) |
| $R^2$ | 0.118 | 0.121 | 0.121 | 0.120 | 0.119 | 0.119 | 0.119 | 0.119 |
| Observations | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 | 1260 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.
[a] $\mu$ is defined as the difference between public and private donations. The binary variable Image Motivation (IM) denotes whether the difference between public and private donations is greater or equal than a threshold, which varies by column.
[b] Models controls by an extended version of the midwife characteristics (reduced plus whether midwife's ethnicity is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the area where she attended pre-school, and whether the midwife lived during primary school that was more rural than the place where currently living now), the inclusion of state dummies (state1-state36), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives, and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

Table A.6: Monetary Incentives and Social Acceptability

| | Model A[a] | Model B[a] | Model A[b] | Model B[b] |
|---|---|---|---|---|
| Monetary Only | 1.674** | 1.299* | 1.315** | 1.122** |
| | (0.760) | (0.662) | (0.619) | (0.544) |
| Non-Monetary Only | 0.875 | 0.836 | 0.596 | 0.439 |
| | (0.834) | (0.558) | (0.608) | (0.399) |
| Both | 1.406* | 0.923 | 1.306** | 0.838* |
| | (0.764) | (0.559) | (0.541) | (0.446) |
| $R^2$ | 0.0986 | 0.0652 | 0.220 | 0.177 |
| Observations | 1269 | 1269 | 1269 | 1269 |

Note: Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Source: SURE-P DATA.

[a] The dependent variable is measured as "Earliest month at which it is socially or very socially acceptable to quit the job".

[b] The dependent variable is measured as "Earliest month at which it is very socially acceptable to quit the job".

[c] Model A controls for a reduced version of the midwife characteristics (age, age squared, number of years that midwife has been living in the community and its square term, tenure in the PHC in months and a dummy year, whether the midwife lived in a rural area, number of years working as a midwife (in month) and its square, MBI emotional exhaustion raw score, MBI depersonalisation raw score, and MBI accomplishment raw score), WDC and Facility characteristics (number of midwives working in the PHC at baseline, incidence, number of equipment which are present and working and whether the PHC f was in the quality enhancement program), and socio-economic characteristics (percentage of women who achieved primary education grade 6 or higher per cluster, cluster average asset ownership composite index (a la Anderson 2008), one binary variable for CCT-treatment clusters which were part of an experimental evaluation, another binary variables for cluster in which the CCT was implemented non-experimentally), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.

[d] Model B controls for an extended version of the midwife characteristics (reduced plus whether midwife's ethnicity is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the catchment area of the PHC where she works, whether midwife's religion is the same as of the area where she attended pre-school, and whether the midwife lived during primary school that was more rural than the place where currently living now), the inclusion of state dummies (state1-state36), and the interaction between randomization strata and a quadratic polynomial on the number of recently hired midwives.