Journal of Economic Behavior and Organization 176 (2020) 489–511

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Conspicuous monitoring and remote work☆

Nathan Jensen[a], Elizabeth Lyons[b,*], Eddy Chebelyon[a,c], Ronan Le Bras[d,e], Carla Gomes[d]

[a] International Livestock Research Institute, Nairobi, Kenya
[b] School of Global Policy & Strategy, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[c] Fraym, 7900 Westpark Dr. McLean, VA 22102, USA
[d] Department of Computer Science, Cornell University, 353 Bill and Melinda Gates Hall, Ithaca, NY 14830, USA
[e] Allen Institute for Artificial Intelligence, 615 Westlake Ave N. Seattle, WA 98109, USA

ARTICLE INFO

Article history:
Received 7 October 2019
Revised 16 April 2020
Accepted 15 May 2020
Available online 17 June 2020

JEL classification:
J24
M54
D83
O13

Keywords:
Moral hazard
Monitoring
Remote work
Field experiment

ABSTRACT

Credible monitoring of remote workers presents unique challenges that may reduce the benefits of formal organization for their management. We consider whether increasing the salience of monitor productivity without changing incentive contracts or monitoring technology leads to changes in remote worker performance. Results from a field experiment run among multi-dimensional task workers in Kenya demonstrate that increasing the visibility of monitor activity improves performance on task dimensions not being directly paid for. Our evidence is consistent with the importance of conspicuous monitoring when managers and workers are not co-located.

1. Introduction

The importance and difficulty of monitoring workers in order to properly reward and punish them is an important motivation for the existence of organizations (Alchian and Demsetz, 1972).[1] As remote work becomes more common

☆ The authors are grateful to Chris Barrett, Andrew Mude, and conference and seminar participants at the International Livestock Research Institute, NYU's Stern School of Business, UC Merced's Department of Economics, UCSD's Rady School of Management, and the 4th International Conference on Computational Sustainability for valuable feedback. Rich Bernstein, Oscar Naibei, and Yexiang Xue provided excellent research assistance. We gratefully acknowledge funding and resource support from NSF Expeditions CompSustNet: Expanding the Horizons of Computational Sustainability, Award CCF-1522054; NSF (0832782, 1059284, 1522054); ARO grant W911-NF-14-1-0498; The Center on Global Transformation at UC San Diego; the Policy Design and Evaluation Lab at UC San Diego; the Atkinson Center for a Sustainable Future's Academic Venture Fund, and Australian Aid through the AusAID Development Research Awards Scheme. The views expressed in the publication are those of the authors and not necessarily those of their funders. The authors thank the International Livestock Research Institute for generous hospitality. This study is registered in the AEA RCT Registry and the unique identifying number is: "AEARCTR-0001848".

* Corresponding author.
E-mail addresses: n.jensen@cornell.edu (N. Jensen), lizlyons@ucsd.edu (E. Lyons), e.chebelyon@fraym.io (E. Chebelyon), ronanlb@allenai.org (R.L. Bras), cmw84@cornell.edu (C. Gomes).

[1] Alchian and Demsetz (1972) argue there are situations in which the market mechanism will not appropriately rewards agents for their efforts, for instance in the case of joint production. In those situations, a residual rights holder can better match rewards to effort by monitoring inputs.

https://doi.org/10.1016/j.jebo.2020.05.010
0167-2681/

(e.g. Bloom et al., 2015) and as monitoring technology advances, firms are increasingly using IT-based solutions to monitor worker inputs and outputs (Bernstein, 2017; Bresnahan et al., 2002). However, some task types may be difficult to track through IT programs. For example, monitoring remote work not performed online is logistically challenging. As a result, even when optimal inputs are definable, input-based incentive pay may not be optimal because inputs cannot be observed or verified (Prendergast, 2002). Consistent with this, Holmstrom and Milgrom (1991) suggest that employees who work from home should have their pay more closely linked to outputs than those who work in the office. However, when output is also costly to measure accurately, output-based pay on its own may be insufficient for optimizing worker performance.[2]

In general, linking pay to measurable dimensions of performance is important for ensuring workers are accurately rewarded. As a result, economic theory suggests that salaries, or input-based pay, may lead to better overall performance when at least some dimensions of output are hard to measure well. Conversely, when workers have information that managers do not have and when worker inputs cannot be verified, output-based should dominate.[3] However, ensuring good performance on difficult to measure task dimensions may also be essential for an organization. Simply communicating that these dimensions matter without accompanying signals that they matter enough for managers to monitor or pay for them may not be sufficient for workers to invest effort in them. Moreover, while much of the work on principal-agent theory assumes the principal does not have private information (Maskin and Tirole, 1990), even if managers claim they will monitor these dimensions that are not directly paid for, the extent to which they are able and willing to effectively do so is typically unknown to workers ex-ante.

In this paper, we test a possible solution to overcoming some of the difficulties associated with managing workers whose inputs and outputs are costly to measure. In particular, we examine whether low-cost increases in the visibility of monitoring without accompanied changes in monitoring technologies or in payment contracts affects remote worker performance. To test this, we designed and implemented a field experiment on a population of workers who are particularly difficult to monitor.

Our field experiment involves the random assignment of a control and two treatments – a monitoring treatment and a check-in treatment – to workers tasked with collecting, classifying, and transmitting data on rangeland conditions in rural areas of Northern Kenya. These workers submit large quantities of data based on information not available to managers, making the quantity of worker output easy to monitor but quality difficult to assess. In both the monitoring and check-in treatment groups, workers received a phone call from their direct manager, or supervisor, once every five days during the treatment period. In the check-in group, the manager told workers how much data had been received the previous day, and how much of that data had a specific characteristic unrelated to the quality of work performed.[4] The manager did not give workers any evaluation-based feedback on the quality or quantity of data received. In the monitoring group, the manager provided the same information given to workers in the check-in group, and also told the worker how much of the received data was of poor quality on two dimensions. The control and two treatment groups allow us to examine the impact of increasing the salience of management and the added effects of signaling to workers that their work quality is being actively monitored and evaluated. In no case was payment or the conditions for termination adjusted in response to the treatments, and all workers were paid according to the same quantity-based protocols regardless of their treatment status.[5]

Given the existing evidence on incentive design, worker observability, and performance, the expected impact of our intervention is theoretically ambiguous. First, the intervention does not change the dimension of output workers are being paid for or the conditions under which they can be terminated. As a result, workers may not be sufficiently incentivized to change their behavior in response to the change in managerial activity.[6] However, evidence from IT-based monitoring interventions among restaurant workers (Pierce et al., 2015) and health among workers (Staats et al., 2016) demonstrate that changes in monitoring technologies without accompanied changes in incentive contracts can improve worker performance on the task dimension targeted by the monitoring intervention. In a setting very similar to ours, Kelley et al. (2018) find a positive impact on worker performance and firm profits from the introduction of a new monitoring technology to track the performance of minibus drivers. Although our intervention does not involve a change in monitoring technology, this

[2] Output may be difficult to evaluate, for instance, in non-profit programs where impact evaluations are not feasible (Weisbrod, 1989), tasks that require hard to verify information, or tasks that require subjective performance assessments (Gibbons, 1998).

[3] Much of this work is described in more detail in Gibbons (1998). Findings from these theoretical analyses demonstrate important trade-offs between input and output-based pay. For instance, input-based pay reduces difficulties and costs associated with monitoring output but also lowers sorting efficiency (Lazear, 1986). Consistent with input-based pay reducing sorting efficiency, higher uncertainty about worker effort is associated with an increased likelihood of output-based payment schemes (Prendergast, 2002). Similarly, output-based pay may dominate when managers can feasibly measure a type of performance that contributes to the organization's objectives and when workers have access to information about the task that managers do not have (Baker, 1992). Empirical evidence on the relationship between monitoring and payment schemes largely support these theories (e.g. Courty and Marschke, 2004; Cragg, 1997).

[4] Specifically, the manager reminds workers how many of the data points they reported to have grass in them. As we discuss in Section 2, this dimension of output is easily observable to the manager, however whether or not workers *accurately* classify data points as having grass is not communicated to workers in this treatment group. Thus, the check-in treatment does not include a meaningful signal of output quality monitoring activity.

[5] This allows for the possibility that quality measures of performance may be imperfect and, therefore, that directly tying pay to subjectively assessed quality may be difficult to contract for (Baker et al., 1994).

[6] For instance, as proposed by Holmstrom and Milgrom (1991), evidence from multidimensional task settings demonstrate that workers respond to task dimensions they are directly paid for at the expense of those they are not even when there is a credible threat of punishment from poor performance on these other task dimensions (e.g. Hong et al., 2013; Larkin and Pierce, 2015).

evidence does suggest that increasing the saliency of monitoring on its own may improve performance on task dimensions not being directly paid for.

A second reason for the theoretical ambiguity associated with our intervention is the uncertainty associated with how workers will interpret the increased contact with their manager. Reminding workers that monitors are paying attention to the quality of their performance without paying them directly for it may positively impact performance if workers believe they can be credibly terminated based on quality assessments (Pierce et al., 2015), or if it leads workers to feel more valued or more important for the organization's success (Kim and Hamner, 1976; Rosen et al., 2006; Tziner and Latham, 1989). Alternatively, consistent with the evidence presented in De Jong and Dirks (2012) and Frey (1993), monitoring may reduce worker performance if workers interpret it as manager distrust or, relatedly, if it crowds out workers' intrinsic task motivations (Ranganathan and Benson, 2016). For instance, de Rochambeau (2017) finds that increased monitoring among Liberian truckers, a setting with weak employment contract enforcement, has negative impacts on the performance of those who perform well prior to the increase. Moreover, even if increased monitoring is not interpreted as managerial mistrust, given that this signal cannot be communicated in-person when managing remote workers, it may not be sufficient to alter worker performance.[7]

Findings from our experiment are consistent with an increase in the observability of monitoring activity improving worker performance. In particular, we find that the monitoring treatment improves worker performance on task dimensions not being directly paid for and even on dimensions not mentioned during the treatment calls, and that the treatments had no significant effect on performance of task dimensions workers are directly paid for. We find that these effects are significantly larger than the effects of an increase in managerial visibility, via check-in phone calls, alone. We present evidence that the improvement in performance is driven by increased job satisfaction among monitored workers, and not by differential learning or changes in reciprocity considerations.

These findings demonstrate that small changes in the saliency of output quality monitoring increases performance even when worker pay is unaffected. Conversely, our evidence shows that remote workers may shirk when they are not receiving signals that they are being monitored on dimensions of performance that employers have stated are important. Combined, our findings demonstrate that, in situations where fully observing and measuring inputs and outputs is not possible, increasing the visibility of manager productivity, even along a subset of activities, may overcome the deficiencies associated with performance-based pay.

The results of our study are consistent with existing research that has also found evidence of a relationship between worker perceptions about managerial productivity and worker performance. For instance, Nagin et al. (2002) provide experimental evidence in support of the rational cheater model among on site workers when employees have a less positive attitude towards their employers, but that those who feel positively about their employer do not increase shirking in response to a fall in monitoring. Unlike our setting, Nagin et al. (2002) do not analyze multiple performance dimensions, nor do they analyze opportunistic behavior among remote workers. Moreover, we find that employee attitudes towards managers may themselves be a function of monitor activity. Al-Ubaydli et al. (2015) do analyze multiple performance dimensions, and test whether uncertainty about the monitor's productivity can overcome challenges associated with providing incentives in tasks with quality and quantity performance requirements. The authors provide theoretical and empirical evidence that with two-sided asymmetric information about worker effort and manager ability, quantity-based pay can lead to better quantity and quality than fixed wages because it acts as a signal that the manager is productive. In our setting, we hold quantity-based pay constant and vary the signal of manager productivity by adjusting the extent to which the manager communicates with workers about their quality. As in Al-Ubaydli et al. (2015), our findings are consistent with signals of managerial productivity improving worker performance.[8] By providing causal field evidence on the impact of increased monitor visibility on remote and multidimensional task worker productivity, holding incentive contracts constant, our study contributes novel evidence that low-cost increases in the visibility of manager productivity can have significant impacts on worker performance even along task dimensions managers are not observed to be monitoring. Moreover, we provide evidence that higher observed monitoring productivity may be improving performance by improving job satisfaction.

More generally, our study contributes to the growing managerial literature the impact of increased worker observability through technological advancements. As Anteby and Chan (2018) highlight, it is increasingly feasible to collect information on almost all aspects of worker performance. However, given the high costs of classifying much of this information into data that can be evaluated by managers[9], and evidence of the negative impacts of increased surveillance on worker performance (Bernstein, 2012), it is unclear how this information should be used by managers. Our study suggests that a promising use of big data for labor management is for remote worker supervision where workers may be less aware of the extent to

[7] Telephone based communication may not be as effective as face-to-face communication because non-verbal cues that may be important for mutual understanding of what is being communicated are not conveyed over the phone (Warkentin et al., 1997). For instance, trust may be lower in virtual teams than in co-located teams (Jarvenpaa and Leidner, 1999) which could hinder monitors' abilities to convey credible threats to remote workers. More generally, remote worker supervision is associated with lower worker performance, particularly when workers are not co-located with team members as in our setting (Bonet and Salvador, 2017).

[8] Also related to our paper, Lu (2012) finds evidence that increasing monitoring on some dimensions of quality reduces performance on the other dimensions in nursing homes where consumers make purchasing decisions based on the monitored quality dimensions. This is distinct from our setting where we are altering monitoring activity on dimensions of performance that workers are not receiving direct incentives for.

[9] For instance, to achieve this, it may be necessary to hire data scientists from a small supply of available talent (McAfee et al., 2012).

Fig. 1. Study region and training sites. *Notes:* This map displays the location of the training sites in Kenya and in relation to one another.

which they are being monitored. Furthermore, they suggest that selective use of fine-grained performance data in occasional performance reviews balances the need to ensure workers believe their performance is valued and accounted for with the importance of preserving worker autonomy and avoiding unnecessary disruptions (Bernstein, 2017), and that some active monitoring can, in fact, increase job satisfaction.

In addition to contributing to the literature on organizations, management, and labor economics, we contribute to the growing literature on management and economic development (e.g. Bloom et al., 2016). In particular, firms are increasingly recognizing the profit potential in more rural areas of developing countries (Neuwirth, 2014; Reardon et al., 2003) but poor infrastructure has made it difficult for firms to establish distribution channels to these regions (e.g. Dihel, 2011). Employing teams of remote workers who reside in these regions may help to overcome these hurdles. However, these work arrangements introduce significant managerial challenges (Bilal et al., 2011). Low-cost solutions for maintaining quality and quantity of output by remote workers have significant implications for private sector development in emerging markets.

This paper proceeds as follows. Section 2 presents the experiment design; Section 3 describes and summarizes the data; Section 4 presents the empirical results from the experiment and an analysis of mechanisms; and Section 5 summarizes and concludes.

2. Experimental design

To test whether active and visible monitoring of output changes remote worker performance, we employed a field experiment among remote workers in rural Kenya. In this section of the paper, we describe the population of workers in our sample, our treatment groups, and the implementation of our treatments.

2.1. Study setting and population

We ran our experiment on 113 workers who were hired to collect and transmit information on rangeland conditions in rural semiarid areas of Northern Kenya over a 149 day period between March and August 2015. Fig. 1 demonstrates the 150 km by 150 km study region which covered parts of Samburu County, Isiolo County, and Laikipia County. The rangeland data collection was part of a collaborative effort between the International Livestock Research Institute in Nairobi and Cornell University in Ithaca, New York to test the viability of information crowdsourcing as a means for improving resource

Fig. 2. Example rangeland photos. *Notes:* These are photos of rangelands taken in the study region. They were used for the training of workers in the study.

allocation among pastoralist communities.[10] Given the difficulties associated with finding labor to work in very remote regions and the knowledge required to classify local rangelands, workers were hired from the population of pastoralists active in the region.[11]

The majority of the working-aged population in poor countries earn their living through rural agricultural production (The World Bank, 2019), and there is a substantial amount of international aid dedicated to improving rural poor agricultural productivity.[12] Despite this, high quality information on production decisions and environmental conditions in these settings has historically been unavailable. Crowd sourcing of this information from local experts who have a deep understanding of the land is increasingly being employed to address this lack of information.[13] While this approach is potentially promising, a serious concern with it is the difficulty associated with certifying and monitoring the data collectors and, thus, ensuring their output is accurate and useful.

In order to collect and transmit information on rangeland conditions, pastoralists were supplied with GPS enabled smartphones with moderate resolution cameras.[14] A crowdsourcing mobile application was developed for the purpose of this job, and the workers submitted all their data through the application. To achieve a single completed submission, workers were required to take a photo of a rangeland and indicate whether the rangeland in the photo includes any grass, trees, or bushes, and, if so, whether each vegetation type is green or brown. In addition, workers were required to indicate the carrying capacity of the rangeland for cattle.[15] Examples of rangeland photos taken in the study region and used as examples of good quality photos in the training these workers are presented in Fig. 2. Workers were paid between $0.05 and $0.40 per submission for up to ten submissions per day.[16] To ensure that they did not submit multiple photos of the same rangeland within a short time period, photos had to be submitted one hour apart. Moreover, to ensure rangelands would be visible in the photos, submissions had to be recorded between 7 am and 6 pm. Submissions that did not meet these qualifications were not paid for. The timing restrictions combined with the ten submission maximum eligible for pay per day ensure that meeting the quantity quota does not require workers to sacrifice quality. In particular, even if completed with a lot of effort and thought, the survey takes well under an hour to complete. With one photo per hour, workers have plenty of time to allocate to both photo and survey quality, and leisure or other pursuits while still submitting the maximum output they can be paid for.

[10] See https://www.udiscover.it/applications/pastoralism/ for more information on the motivation for the workers' tasks.

[11] For further information on pastoralism in Northern Kenya, see for instance McPeak and Barrett (2001).

[12] For instance, in 2018 USAID alone spent 906M USD on agricultural development programs https://explorer.usaid.gov/agencies).

[13] For example, https://www.premise.com/ provides a platform the link data collectors on the ground in rural agricultural settings and in others with aid agencies, governments, and firms seeking information that would otherwise be almost impossible to gather.

[14] We tested multiple smartphones based on their ability to take and geo-locate photos while having a long battery life.

[15] Some pastoralists hired for this work are not literate or fluent in English, and some are not literate in any language. To ensure literacy was not required to complete the task, workers completed each classification step by selecting images that corresponded to their responses.

[16] Payment varied with the location photos were taken in an effort to increase data collection from more remote locations.

Workers received three days of intensive training on the use of the smartphone, the application, and the task at one of the five training sites shown in Fig. 1. During this training, workers were repeatedly told that a good quality submission is one with accurate classifications, and landscape images that are clear and capture as much of the rangeland as possible. They were also told that if they were found to be shirking on the quality of their submissions, their employment would be terminated and their smartphones would be repossessed. Moreover, during training workers were shown how submissions could be downloaded from the server and linked to classification surveys to evaluate the accuracy of submission classifications. Thus, while the accuracy of submissions and quality of photos were not directly paid for, these task dimensions were explicitly incentivized in the employment contract through the threat of job loss. Workers were repeatedly quizzed on their understanding of how to complete a high quality submission, and on their understanding of their employment contract. By the end of training, all workers demonstrated a strong understanding of their job expectations. No workers were fired during the employment period.

There are several dimensions of data submission quality that are relatively easy to verify, and several that are quite difficult. In particular, the location of the photo, the time it was taken, whether it had been previously submitted, and how photos were classified are automatically recorded through the application and easy to verify as a result. Location and time of the photo are particularly important to verify because payment is conditional on these characteristics. In contrast, the accuracy of the classifications made and the quality of the photos are difficult to verify because of the large quantity of data submitted.[17] Workers may have an incentive to mis-classify photos to reduce the time it takes to submit each one if choosing random options on each screen in the application is faster than choosing the correct option. Similarly, it may be easier to submit quickly taken, poor quality photos. By saving time on image classification or photography workers can increase time spent on leisure or on other income-earning activities.[18] In addition, if they believe that aid to the region would be affected by the crowdsourcing effort,[19] then they may have an incentive to classify photos as have worse rangeland conditions than in reality. In fact, our data suggests this may be occurring; 70% of the total number of instances in which grass was misclassified were due to workers under-reporting its presence.

2.2. Experimental treatments

We introduced two managerial treatments within the rangeland data collection effort. Workers assigned to the first treatment, which we will hereafter refer to as the “check-in” treatment, received a call from their supervisor every five days. During the call, the supervisor told each worker how many submissions he[20] had made the previous day, and how many of those submissions were classified as having grass in them. This was done as an indication that the supervisor was able to observe both the quantity of submissions and their content. The supervisor did not give workers any evaluation-based feedback on the quality or quantity of data received, and, importantly, did not tell workers whether the photos were correctly specified as having grass in them. Thus, the check-in treatment did not provide any new information about dimensions of performance valued by the employer. The reason that we included a statement about a dimension of output unrelated to the quality of worker performance was to ensure the calls were not perceived as completely redundant.[21,22,23]

Workers assigned the second treatment, which we will hereafter refer to as the “monitoring” treatment, also received a call from their supervisor every five days. The beginning of the call was identical to the call in the check-in treatment. However, workers in this treatment group were also told which submissions from the prior day they had correctly and incorrectly classified the presence of grass in the photo. In addition, the manager told workers how many submissions from the prior day included poor quality photos and were reminded that photos should be taken during the day, not be blurry, and capture a wide scene. The precise scripts the manager read workers in the respective treatments are as follows:

[17] At the time of the study and of writing, image classification algorithms were not sufficiently advanced for us to automatically verify the accuracy of submissions.

[18] Worker incentives to shirk on the quality of their submissions in our setting is theoretically consistent with the incentives for restaurant worker theft described in Pierce et al. (2015). In particular, workers in our setting can earn income through productivity in image submissions or in other activities. If the expected cost of submitting low quality images is sufficiently low, for instance because workers do not expect to get caught, they may optimally shirk on classification quality in order to increase time on other income-earning activities.

[19] During training, pastoralists were told that one of the objectives for the crowdsourcing effort was to improve the classification of rangelands in order to improve emergency aid targeting among other things.

[20] All workers in our sample are male.

[21] Communication about the total number of submissions was redundant because workers were receiving payment for this performance measure and, thus, knew the supervisor was aware of how many photos they were submitting. We were concerned that if the check-in calls only provided redundant information, they would have a large negative effect on performance by indicating the supervisor was very low quality.

[22] Although workers were trained on the importance of accurate submission classifications and not on the importance of submitting particular types of rangeland, it is possible that the mention of submissions with grass in them in the check-in treatment altered workers' understanding of the employer's grass classification desires. If the check-in calls did have this effect on performance, it would be difficult to interpret them as a placebo treatment. We verify that the check-in treatment did not change the quantity of submissions classified as having grass in them. Table A.9 demonstrates no change in the frequency of grass reporting by workers in the check-in treatment. Moreover, we find no change in the accuracy of grass reporting in response to the check-in treatment (Table 5). Combined, these findings are not consistent with the mention of grass in the check-in treatment changing the types of photos workers submit, or how they classify the photos.

[23] To ensure that the treatments were administered as expected, the local manager recorded and transmitted data on the calls he made, and any questions workers had during those calls.

Check-in treatment: Our records show that yesterday you completed and submitted [xx] surveys and that in [yy] of those surveys you indicated that there was grass.

Monitoring treatment: Our records show that yesterday you submitted [xx] surveys and that in [yy] of those surveys you indicated that there was grass. We agree with your grass categorization in [z1] cases but disagree in [z2] cases. Do you remember why you might have said there was no grass when there was grass or some grass when there was none in the photo? Our records also show that there were [z3] cases in which the photo was of very poor quality. Please remember that photos must be taken during the day, not be blurry, and you must stand back from objects so that the photo captures a wide scene.

The manager was instructed not to give any additional feedback or comments on the workers' performance or submissions and to make notes of all questions and comments from the workers during these calls.

It is important to note that what we are calling our monitoring treatment includes information about what the manager is observing about worker output, and how the manager is evaluating worker output. The reason for this is that simply stating what the manager is observing as in the check-in treatment may not be a credible signal of managerial productivity. More specifically, under the check-in condition, workers may not infer that their output is being observed for the purposes of evaluation, which is the objective of managerial monitoring (e.g. Katz, 1986; Lazear, 1991). While conspicuous monitoring when workers and managers are co-located may not require managers to consistently state their evaluation of workers, in the context of remote work this is harder to achieve. Therefore, our findings on the effect of our monitoring treatment should be interpreted as the effects of communicating both that managers are both observing performance dimensions not being directly paid for and providing feedback on how well workers performed on those dimensions.

2.3. Study implementation

The check-in and monitoring treatments were each randomly assigned to 34 workers in the study population, and the remaining 45 workers did not receive any phone calls from the local manager. Treatments were randomly assigned within each training site to ensure that each site has workers in all three groups. The manager called all workers in the treatment groups in a single site per day. These calls began 43 days into the study period. To test whether the treatments continued to have effects after the calls stopped and help us examine treatment effect mechanisms, we phased the calls out gradually. Specifically, we dropped 25% of the treatment group from the call list at a time with the first 25% being dropped 52 days after the start of the treatments and each subsequent 25% dropped after 15 days. All calls stopped 15 days before the end of the study period.

At the beginning of the study period, workers were surveyed by their local manager. This questionnaire asked about educational and work backgrounds, demographics, and normal phone use. Workers were told that their activities would be used to study the viability of crowdsourcing for improving information on range land conditions and related topics, but did not know that we were studying questions related to worker management or that managerial interventions were being randomly assigned, or assigned to a subset of workers. As part of their participation in the study evaluating the crowdsourcing method, all workers signed a consent form that stated the anonymized submissions could be shared with and evaluated by people other than their supervisor and those involved in training them.

3. Data and estimation strategy

3.1. Data description

Our study includes data from 149 days of worker activity from March to August in 2015 and our experimental treatments were implemented from April 24 to July 26 of that year. Our sample includes the population of workers hired for the rangeland crowdsourcing project. Of the workers in the treatment groups, 14 received the treatment over the course of a 49 day period, 17 received treatment over 64 days, 15 received treatment over 79 days, and 22 received treatment over 94 days. We have before and after treatment observations for all workers.

Data on worker activity was collected from the server workers sent their survey submissions to. This includes the number of submissions each worker made each day, the location and time the photos included with the submissions were taken at, and each rangeland classification corresponding to each photo. We collected data on the quality of submitted photos and the accuracy of accompanying classifications for a random subset of submissions. Specifically, we posted a random subset of anonymized submissions stratified by treatment group on a Mechanical Turk project and had Turkers classify the photos and indicate whether they met certain quality standards.[24] We asked Turkers whether each photo includes any grass, whether it is blurry, back-lit, taken at a bad angle, or of poor quality for any other reason. We chose these dimensions to ensure we had data on the dimensions of quality mentioned in the monitoring phone call as well as some dimensions that were not mentioned in the call to check whether workers sacrificed performance on dimensions not being visibly monitored for those that were. We did not ask Turkers to classify the presence of trees or shrubs because in pilot testing on Mechanical

[24] The monitoring treatment required that submissions from each day prior to each phone call be reviewed by Turkers in order for the supervisor to provide workers with feedback. We include these reviews in our measure of quality. Thus, more submissions were evaluated for quality in the monitoring group than in the others.

Table 1
Worker summary statistics.

	Mean	Std. dev.	Min	Max	No. of obs.
Panel A: Worker characteristics					
Age	22.369	3.519	18	35	111
Male	1.000	0.000	1	1	111
Current student	0.099	0.300	0	1	111
Highest level of education	2.09	0.837	0	5	111
Years of herding experience	13.324	5.534	2	27	111
Average number of calls per day	9.919	7.932	1	50	111
Average number of SMS' sent per day	40.874	58.994	0	300	111
Panel B: Output characteristics by worker					
Average number of submissions per day	9.361	2.168	4.013	14.535	113
High quality submissions per day	4.868	1.381	1.619	7.859	113
Proportion of days with at least ten submissions	0.652	0.241	0.028	0.967	113
Worker left job	0.097	0.298	0	1	113
Grass accuracy	0.613	0.487	0	1	24,969
Night time photo	0.005	0.073	0	1	90,925
Poor quality monitored (blurry or bad angle)	0.107	0.309	0	1	32,562
Poor quality not monitored (backlit or bad quality other)	0.204	0.403	0	1	32,562

Turk we learned that these classifications were generally very inaccurate, perhaps due to difficulties distinguishing shrubs from trees in photos.

A total of 107,286 submissions were made by the workers. Of these, 38,013 submissions were reviewed by Turkers (32,562 of which were submitted before or during treatment), and, of these, 28,572 were reviewed by at least 2 Turkers. The variables we generated from these reviews are: poor quality monitored, which captures the dimensions of photo quality discussed in the monitoring treatment calls and is equal to one if any submission reviewer indicated the photo is blurry or taken at a bad angle and zero otherwise; poor quality not monitored, which captures dimensions of photo quality not discussed in the monitoring treatment calls and is equal to one if any submission reviewer indicated the photo is back lit or of poor quality for some other unspecified reason and zero otherwise; grass accuracy, equal to one if more than 50% of Turkers agree with how the presence of grass was classified by the pastoralist who submitted the survey and zero otherwise. We restrict this last measure to submissions that were reviewed by at least 2 Turkers, however, our results are robust to including the full set of Turker reviewed submissions. Measures of picture quality and accuracy of grass reporting are also used to generate a measure of the number of high quality submissions made per day; a measure that is a daily count of the number of submissions that were not flagged for being of poor quality along any dimension among those photos only reviewed by one Turker and that were not flagged for being of poor quality and accurately classified the presence of grass among those photos reviewed by at least two Turkers. To account for differences in the number of photos reviewed by external reviewers across treatment and control groups, we divide this measure by the worker-day proportion of submissions reviewed and set the measure equal to zero on days workers did not make any submissions.

We also collected data from worker characteristic and demographic surveys that all but two workers filled out. This data includes information on worker age, education, job experience, and phone use norms.

Table 1 presents summary statistics on worker activity, characteristics, and submission quality. Panel A reports summary statistics on worker characteristics, and Panel B reports summary statistics on worker output. As Panel A demonstrates, all workers are male and are relatively young with an average age of 22 and a maximum age of 35. In addition, workers are active phone users averaging about 10 calls and 40 text messages per day[25], and have extensive experience herding animals which is a proxy for their familiarity with assessing the quality of local rangelands and with the geography of the region.

Panel B summarizes characteristics of output that are most relevant for evaluating our treatment effects. The first five variables summarized in Panel B, specifically the average number of submissions per day, the average number of high quality submissions per day, the proportion of days with at least ten submissions, and whether the worker quit or left the job early are at the worker level and demonstrate that workers were very active during the study period. In particular, workers often submitted approximately the maximum number of pictures they would be paid for. The remaining variables summarized are at the submission level. Of the just over 107,00 submissions received, 16,075 were submitted during the post-treatment study period leaving us with 90,925 for our main analysis that compares the pre-treatment period to the treatment period. Of these 90,925 submissions, about 61% accurately classified the presence of grass. While this appears quite low, as the rangeland photos presented in Fig. 2 demonstrate, it can be quite difficult to distinguish between shrubbery and grass, and, when grass is brown, between dirt and grass. This is also why we required at least that two Turkers evaluate the grass in each photo. In addition, less than 1% of photos were taken outside of the permissible hours. Among the 32,562

[25] Including WhatsApp, Facebook, and Viber messages.

Table 2
Worker characteristics by treatment and control groups.

	Control	Check-in	Monitoring	*p*-value[+]
Characteristics				
Age	22.349	22.000	22.765	0.673
	(0.3993)	(2.697)	(3.660)	
Current student	0.116	0.118	0.059	0.647
	(0.324)	(0.327)	(0.239)	
Highest level of education	2.279	2.00	1.941	0.161
	(0.984)	(0.550)	(0.851)	
Years of herding experience	13.744	13.000	13.118	0.817
	(5.774)	(6.218)	(4.538)	
Average number of calls per day	10.930	10.412	8.147	0.285
	(8.795)	(8.937)	(5.153)	
Average number of SMS' sent per day	40.209	42.647	39.942	0.978
	(58.322)	(60.807)	(10.247)	
N	45	34	34	
Pre-treatment performance				
Output characteristics by worker-day				
Average number of	7.942	7.376	8.057	0.000***
Submissions per day	(0.113)	(0.132)	(0.128)	
High quality	3.680	3.382	3.919	0.059*
Submissions per day	(0.098)	(0.100)	(0.110)	
Proportion of days with	0.525	0.509	0.557	0.059*
At least ten submissions	(0.013)	(0.015)	(0.014)	
N[+]	1,571	1,185	1,175	
Output characteristics by worker				
Average number of	9.903	9.420	9.563	0.532
Submissions per day	(0.272)	(0.295)	(0.404)	
High quality	4.436	4.273	4.590	0.534
Submissions per day	(0.185)	(0.168)	(0.211)	
Proportion of days with	0.685	0.680	0.689	0.990
At least ten submissions	(0.033)	(0.040)	(0.042)	
N[+]	45	34	34	
Output characteristics by submission				
Grass accuracy	0.566	0.585	0.599	0.033**
	(0.008)	(0.010)	(0.010)	
Night time photo	0.004	0.002	0.001	0.000***
	(0.001)	(0.000)	(0.000)	
Poor quality monitored	0.124	0.118	0.132	0.257
(blurry or bad angle)	(0.005)	(0.006)	(0.006)	
Poor quality not monitored	0.191	0.212	0.201	0.091*
(Backlit or bad quality other)	(0.006)	(0.008)	(0.007)	
N[+]	39,900	23,128	27,887	

Notes: Standard errors are in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.
[+] Test for equality of three group means using multivariate analysis of variance.

photos reviewed external evaluators, about 20% were classified as bad quality on non-monitored dimensions, and 11% were classified as bad quality on monitored dimensions.

To verify the randomness of worker assignment to treatment groups, we compare average worker characteristics across our treatment and control groups and report these comparisons in Table 2. We do not find any significant differences in these characteristics across groups in pair-wise comparisons of the groups or a test of mean equality across the three groups, which we report the *p*-values from in Table 2. Although random assignment was performed before the study period began and, thus, pre-treatment performance did not enter into our randomization process, Table 2 also checks whether workers in each group differ in their pre-treatment period performance. We find economically small but statistically significant differences in their performance on the quantity of submissions made, high quality submissions per day, the proportion of days with at least 10 submissions made,[26] accuracy of grass reporting, nighttime submissions, and non-monitored quality dimensions across the groups. Given the large number of observations of submissions in each category, it is perhaps not surprising that we see some statistical significance.

However, tests of equality across characteristics when randomization has occurred may not be particularly useful (Bruhn and McKenzie, 2009). As an alternative test of the validity of our random assignment, we analyze whether the joint relationship between worker characteristics and treatment assignment is zero by regressing the worker characteristics presented in

[26] If we compare these outcomes at the worker-level as in Table 1 rather than at the worker-day level, all three are statistically the same across groups.

Table 3
Outcomes by treatment and control groups during treatment.

	Control	Check-in	Monitoring	*p*-value[+]
Performance during treatment				
Output characteristics by worker-day				
Average number of	6.306	5.741	7.301	0.000***
Submissions per day	(0.080)	(0.104)	(0.101)	
High quality	3.356	3.199	4.224	0.000***
Submissions per day	(0.060)	(0.078)	(0.090)	
Proportion of days with	0.410	0.369	0.505	0.000***
At least ten submissions	(0.021)	(0.026)	(0.025)	
N[+]	4,349	2,508	2,523	
Output characteristics by worker				
Average number of	8.486	7.818	9.143	0.323
Submissions per day	(0.547)	(0.717)	(0.509)	
High quality	4.430	4.353	5.398	0.082*
Submissions per day	(0.311)	(0.421)	(0.331)	
Proportion of days with	0.563	0.523	0.639	0.353
At least ten submissions	(0.050)	(0.063)	(0.052)	
N[+]	45	34	34	
Output characteristics by submission				
Grass accuracy	0.631	0.635	0.625	0.620
	(0.006)	(0.009)	(0.006)	
Night time photo	0.008	0.009	0.004	0.000***
	(0.001)	(0.001)	(0.000)	
Poor quality monitored	0.130	0.104	0.071	0.000***
(Blurry or bad angle)	(0.004)	(0.005)	(0.003)	
Poor quality not monitored	0.218	0.228	0.187	0.000***
(Backlit or bad quality other)	(0.004)	(0.006)	(0.004)	
N[+]	28,750	15,305	18,971	

Notes: Standard errors are in parentheses.
* Significant at 10%; ** significant at 5%; *** significant at 1%.
[+] Test for equality of three group means using multivariate analysis of variance.

Table 2 on treatment status and run a test for joint orthogonality. This test, which provides further support for the randomness of worker assignment to treatment groups, is presented in Table A.1.

As a first look at whether our treatments had an impact on performance, we compare mean performance across groups while treatment is ongoing and after it stops respectively. These comparisons are presented in Tables 3 and 4.[27] The means presented in Table 3 demonstrate that while the overall quantity of submissions during the treatment period declines relative to the pre-treatment period across all three groups, the quality weighted quantity of submissions is higher in the monitoring treatment group relative to the pre-treatment period whereas it declined in the other two groups. Moreover, the number of high quality submissions per day is significantly higher in the monitoring group than in the check-in or control groups even at the worker-level of observation.

Consistent with an increase in the quality of submissions in the monitoring group, Tables 3 and 4 demonstrate that both the monitored and non-monitored dimensions of picture quality are significantly higher in the monitored group than the check-in or control groups both during and after treatment, and the likelihood a photo is taken at night is lowest in the monitoring group (though the difference is economically very small and reverses direction in the post-treatment period). Mean grass accuracy is an exception to this overall pattern, and does not appear to be impacted by the monitoring treatment. Overall, these differences are consistent with monitoring improving dimensions not directly rewarded.[28]

To further demonstrate the impact of our monitoring treatment on the number of high quality submissions per day, Fig. 3 presents mean high quality submissions per day by treatments and control across the three study periods. This graph demonstrates that these means were similar in magnitude before treatment began, but that mean worker accurate submissions made per day increased significantly in the monitoring treatment group by about 8%, decreased significantly in the control group by about 4%, and remained statistically the same in the check-in group. High quality submissions continue to

[27] As a reminder, the features of submissions that were mentioned in the monitoring treatment calls were the number of submissions made in a day, the accuracy of grass reporting, the time the photo was taken at, and the blurriness and angle of the photo. The check-in treatment calls mentioned the number of the submissions made in a day, and the number that reported the presence of grass. Importantly, the quantity of submissions, and whether or not they are taken during the day time are both rewarded directly through the pay-for-performance contract.

[28] While the difference in quantity of submissions during treatment suggests monitoring may have increased a dimension of performance directly incentivized, as we described in Section 4, the difference in quantity of submissions across groups is almost fully explained by a difference in the likelihood that workers leave the job before the end of the period.

Table 4
Outcomes by treatment and control groups after treatment.

	Control	Check-in	Monitoring	*p*-value[+]
Output characteristics by worker-day				
Average number of	4.475	6.311	6.063	0.000***
Submissions per day	(0.449)	(0.441)	(0.464)	
High quality	2.040	2.857	3.152	0.000***
Submissions per day	(0.150)	(0.117)	(0.135)	
Proportion of days with	0.307	0.480	0.414	0.000**
At least ten submissions	(0.039)	(0.036)	(0.033)	
N[+]	495	1,152	1,134	
Output characteristics by worker				
Average number of	6.136	6.814	8.723	0.092*
Submissions per day	(0.763)	(0.762)	(0.921)	
High quality	2.812	2.939	4.549	0.013**
Submissions per day	(0.379)	(0.454)	(0.519)	
Proportion of days with	0.428	0.521	0.591	0.215
At least ten submissions	(0.062)	(0.074)	(0.064)	
N[+]	45	34	34	
Output characteristics by submission				
Grass accuracy	0.614	0.569	0.614	0.023**
	(0.020)	(0.012)	(0.013)	
Night time photo	0.009	0.012	0.019	0.000***
	(0.002)	(0.001)	(0.002)	
Poor quality monitored	0.130	0.143	0.114	0.001***
(Blurry or bad angle)	(0.013)	(0.007)	(0.006)	
Poor quality not monitored	0.214	0.240	0.153	0.000***
(Backlit or bad quality other)	(0.016)	(0.009)	(0.007)	
N[+]	2472	7740	7303	

Notes: Standard errors are in parentheses.
* Significant at 10%; ** significant at 5%; *** significant at 1%.
[+] Test for equality of three group means using multivariate analysis of variance.

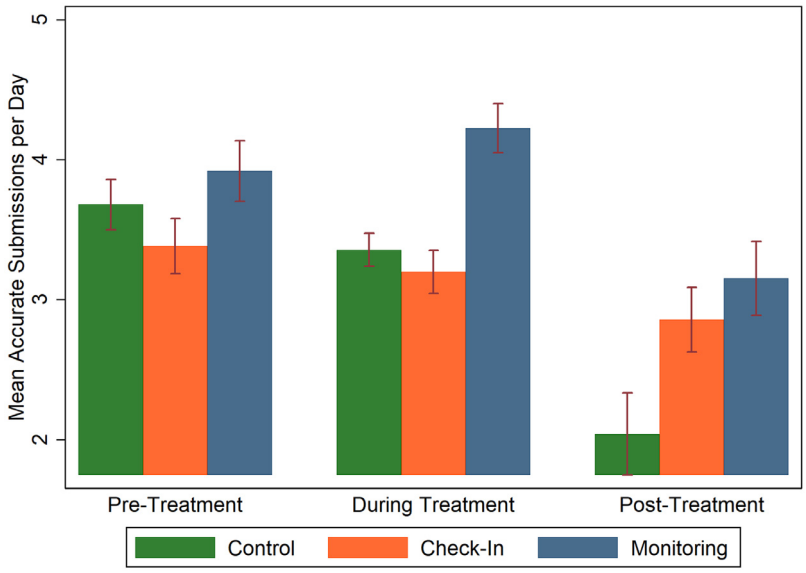

Fig. 3. High quality submissions per day by treatment and period. *Notes:* This graph displays the mean number of high quality submissions made per treatment group across the three periods of observation.

decline in the control group following the treatment period, but interestingly, they also decline significantly in the check-in and monitoring groups.

3.2. Empirical estimation strategy

The mean comparisons discussed above demonstrate that worker performance improved in the monitoring group after treatment began. To examine whether these results hold when we control for differences in locations or individual fixed

effects, we estimate the following equation for outcomes measured at the worker-day level of analysis:

$$Y_{jt} = \alpha + \beta_1 CheckIn_j * Treatment_t + \beta_2 Monitoring_j * Treatment_t + \delta Treatment_t + \theta_1 CheckIn_j + \theta_2 Monitoring_j + Worker_j + \epsilon_{jt}, \quad (1)$$

where Y_{jt} is a measure of average performance of worker j on day t, $CheckIn_j$ is an indicator for whether worker j is in the check-in group, $Monitoring_j$ is an indicator for whether worker j is in the monitoring group, $Treatment_t$ is an indicator for whether day t occurs after treatment began, and $Worker_j$ is a fixed effect for worker j. We also estimate this equation with location fixed effects, treatment group main effects. We cluster standard errors at the worker level. To assess the treatment effect on worker retention, we estimate this equation at the worker level and without the interactions between treatment group and treatment timing.

We similarly estimate the following equation for outcomes at the submission level, where i the submission:

$$Y_i = \alpha + \beta_1 CheckIn_i * Treatment_i + \beta_2 Monitoring_i * Treatment_i + \delta Treatment_i + \theta_1 CheckIn_i + \theta_2 Monitoring_i + \epsilon_i, \quad (2)$$

We should note that our experiment was designed such that multiple outcomes were necessary to analyze in order to evaluate the impacts of our treatments. In particular, to determine how our intervention altered performance, it is important to evaluate changes in the task dimensions mentioned in the monitoring calls and those not included in the calls to verify whether any improvements in performance came at the expense of performance on other dimensions. To address concerns about multiple hypothesis testing (e.g. Savin, 1984), we have taken several precautions. First, our findings all support a single narrative. Moreover, our main findings on the effects of the monitoring treatment reported in Table 5 remain the same when we adjust them using the Bonferroni correction.[29]

4. Empirical results

In this section we report our estimated effects of the check-in and monitoring treatments on worker performance. We first report how performance on these dimensions change during the study period following the beginning of treatment. We then explore possible mechanisms driving the effects of the treatments on worker performance.

4.1. Main results

Table 5 reports the estimated effects of the check-in and monitoring treatments across task dimensions. We analyze how treatment affects the quantity and quality weighted quantity of submissions, the accuracy of grass reporting, whether or not the photos included in submissions were taken at night, whether the photos included in submissions were blurry or taken at a bad angle (both mentioned in the monitoring treatment), whether the photos included in submissions were back-lit or of bad quality for some other reason (not mentioned in the monitoring treatment), and whether workers left the job prematurely. The quantity of submissions, and whether or not grass was reported was also mentioned in the calls workers in the check-in treatment received from their manager.

In both Panels A and B of Table 5, columns 1, and 4 report estimates from regressions with no controls, columns 2, and 5 report estimates from regressions with location fixed effects, and columns 3 and 6 report estimates from regressions with worker fixed effects. In Panel B, columns 7, 8, and 9 follow a similar pattern of gradual control inclusion. In Panel C of Table 5, columns 1 and 4 report estimates from regressions with no controls, columns 2 and 5 report estimates from regressions with treatment site fixed effects, and column 3 reports estimates from regressions with worker fixed effects.[30] Estimates change very little across specifications, so we will focus on the most conservative specifications, which include worker fixed effects, going forward.

Panel A reports how the treatment affected the quantity of worker output and demonstrates that the quantity of submissions fell significantly during the treatment period in all groups, and that this decline is not significantly lower in the monitoring group. Moreover, the coefficient magnitudes in columns 4–6 demonstrate similar patterns as those presented in Fig. 3, in particular, that monitoring increased high quality submissions per day by more than the check-in treatment, but with our controls and fixed effects the significance of the effect of monitoring relative to the control and relative to check-in goes away.

Panel B reports estimates on how the treatments affected the quality of submissions along the dimensions of output quality mentioned in the monitoring treatment calls. Consistent with Table 3, these results show that neither the monitoring nor the check-in treatments significantly affected the accuracy of grass reporting. Coefficients presented in column 6 show the monitoring treatment did not have a significant impact on the frequency of submissions with photos taken at night. Consistent with our treatments not having a significant impact on the quantity of submissions, workers are directly incentivized to not take photos at night[31] and, as such, the tendency to do even in the absence of intervention is very low.

[29] Due to an unfortunate oversight, we did not register our experiment in the AEA registry until after the study was run.

[30] We cannot include worker fixed effects in our analysis of whether workers leave the job because this is a one time and permanent event.

[31] As discussed in Section 2, photos taken after 6 pm are not paid for.

Table 5
Effect of monitoring activity on worker performance.

Panel A: Quantity outcomes						
	Submissions per day			High quality submissions per day		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment period	−1.636***	−1.631***	−1.616***	−0.324	−0.330	−0.317
	(0.461)	(0.461)	(0.464)	(0.294)	(0.293)	(0.288)
Check-in	−0.567	−0.547		−0.298	−0.272	
	(0.533)	(0.553)		(0.283)	(0.311)	
Monitoring	0.115	0.135		0.239	0.275	
	(0.556)	(0.582)		(0.313)	(0.335)	
Check-in ×	0.002	−0.001	0.134	0.141	0.134	0.222
Treatment period	(0.878)	(0.881)	(0.872)	(0.535)	(0.538)	(0.520)
Monitoring ×	0.880	0.886	0.687	0.630	0.638	0.553
Treatment period	(0.733)	(0.741)	(0.729)	(0.510)	(0.512)	(0.504)
Check-in × Treatment=						
Monitoring × Treatment	0.349	0.348	0.547	0.425	0.415	0.582
Location fixed effects	No	Yes	No	No	Yes	No
Worker fixed effects	No	No	Yes	No	No	Yes
Observations	13,310	13,310	13,310	11,581	11,581	11,581
R-squared	0.026	0.068	0.398	0.012	0.027	0.130
Mean dep var	7.395	7.395	7.395	3.573	3.573	3.573

Panel B: Quality outcomes Mentioned in monitoring calls									
	Grass accuracy			Night submission			Blurry or bad angle		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment period	0.065**	0.064**	0.065**	0.004	0.004	0.003	0.006	0.005	0.010
	(0.030)	(0.029)	(0.029)	(0.004)	(0.004)	(0.004)	(0.013)	(0.012)	(0.013)
Check-in	0.019	0.024		−0.002	−0.002		−0.006	−0.008	
	(0.023)	(0.024)		(0.003)	(0.003)		(0.015)	(0.015)	
Monitoring	0.033	0.042		−0.003	−0.003		0.008	0.007	
	(0.023)	(0.026)		(0.002)	(0.003)		(0.013)	(0.013)	
Check-in ×	−0.014	−0.018	−0.030	0.003	0.003	0.003	−0.020	−0.018	−0.025
Treatment period	(0.040)	(0.039)	(0.038)	(0.005)	(0.005)	(0.005)	(0.016)	(0.016)	(0.016)
Monitoring ×	−0.039	−0.040	−0.039	−0.000	0.000	0.001	−0.068***	−0.066***	−0.073***
Treatment period	(0.054)	(0.051)	(0.052)	(0.004)	(0.004)	(0.004)	(0.018)	(0.018)	(0.017)
Check-in × Treatment=									
Monitoring × Treatment	0.641	0.654	0.858	0.311	0.333	0.551	0.005	0.003	0.002
Location fixed effects	No	Yes	No	No	Yes	No	No	Yes	No
Worker fixed effects	No	No	Yes	No	No	Yes	No	No	Yes
Observations	24,969	24,969	24,969	90,925	90,925	90,925	32,562	32,562	32,562
R-squared	0.002	0.008	0.057	0.001	0.003	0.023	0.007	0.008	0.036
Mean dep var	0.613	0.613	0.613	0.00530	0.00530	0.00530	0.107	0.107	0.107

Panel C: Quality outcomes not Mentioned in monitoring calls					
	Backlit or bad quality other			Worker left job	
	(1)	(2)	(3)	(4)	(5)
Treatment period	0.027***	0.028***	0.028***		
	(0.010)	(0.009)	(0.010)		
Check-In	0.022**	0.021*		0.065	0.063
	(0.011)	(0.011)		(0.082)	(0.081)
Monitoring	0.010	0.010		−0.111**	−0.111**
	(0.012)	(0.012)		(0.047)	(0.050)
Check-in ×	−0.011	−0.010	−0.015		
Treatment period	(0.012)	(0.012)	(0.014)		
Monitoring ×	−0.040**	−0.042***	−0.041**		
Treatment period	(0.016)	(0.016)	(0.017)		
Check-in × Treatment=					
Monitoring × Treatment	0.047	0.034	0.128	0.009	0.008
Location fixed effects	No	Yes	No	No	Yes
Worker fixed effects	No	No	Yes	No	No
Observations	32,562	32,562	32,562	113	113
R-squared	0.001	0.002	0.008	0.055	0.078
Mean dep var	0.205	0.205	0.205	0.0973	0.0973

Notes: This table reports estimated effects of treatment on dimensions of the task that were and were not mentioned in the monitoring treatment phone calls. Treatment is equal to 1 during the period that the monitoring or check-in treatments are occurring. Monitoring is equal to one if a worker received monitoring phone calls from his supervisor. Check-In is equal to one if a worker received check-in phone calls from his supervisor. In Panel A, an observation is a worker day. In Panel B and columns 1–3 of Panel C, an observation is a submission. Columns 4 and 5 of Panel C are at the worker level. Robust standard errors are reported in parentheses. High Quality Submissions per Day is the number of submissions not classified as bad quality along any dimension of quality and is restricted to submissions reviewed by external reviewers and weighted by the proportion of photos per participant day that were reviewed by external reviewers. Grass accuracy is equal to one if a submission accurately reports the presence of grass and zero otherwise. Night submission is equal to one if the photo included in the submission was taken at night and zero otherwise. Worker Departure is equal to one if a worker quit or disappeared from the job before the end of the study period and zero otherwise. The Check-in*Treatment=Monitoring*Treatment row reports the p-value of a test of equality between these coefficients. Standard Errors clustered by worker are in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Estimates presented in column 9 demonstrate that, relative to the sample mean, the monitoring treatment led to a more than 70% decrease in the likelihood a submission included a blurry or badly angled photo and that this effect is significantly larger than the effect of check-in on these quality dimensions.

Panel C reports estimates on how the treatments affected outcomes along the dimensions of the task not mentioned in the monitoring treatment calls. Column 3 demonstrates that the monitoring treatment reduced the likelihood that a submission was of poor quality along dimensions not mentioned in the monitoring treatment by about 4 percentage points whereas the check-in treatment did not significantly impact this outcome. Moreover, the monitoring treatment significantly increased worker retention by reducing the likelihood a worker left the job prematurely by 11 percentage points, or more than 100% relative to the sample mean.

Combined, these results demonstrate that increases in visible monitoring improved the quality of worker performance. Perhaps not surprisingly, neither the monitoring nor the check-in interventions significantly improved task dimensions already being directly incentivized. However, monitoring significantly increase quality dimensions of performance relative to both the control group and the check-in group. To put the benefits of the monitoring treatment in financial terms, we generated a measure of the cost effectiveness of our monitoring intervention using the change in high quality submissions per day.[32] We estimate that each dollar spent on this intervention yielded a savings of 1.238USD. Part of the reason the overall quality of submissions did not increase at the same rate as individual quality dimensions is because not all dimensions increased for the same submissions. Thus, this return on investment estimation does not account for submissions that were more high quality than others but not high quality on all dimensions and is thus an underestimate of the intervention's cost effectiveness. In contrast, the cost effectiveness of the check-in treatment is negative with each supervisor call costing about 0.50USD and yielding no change in the quality of submissions.

The one exception to the general trend in our findings is the accuracy of grass classifications, which is not improved by the monitoring treatment. A possible explanation for this could be that the supervisor may not have always provided accurate information in his phone calls, for instance, by inaccurately claiming a worker had misclassified the presence of grass. Given the extent of disagreement about the presence of grass among the external reviewers who also only had photos to go off, even if the supervisor was being as vigilant as possible, occasional inaccuracies in the phone calls are likely. Importantly, this dimension of output is much harder to verify than the others that were included in the monitoring treatment. To check whether imperfect monitoring is related to the decrease in grass accuracy, we re-run our analysis of the effects of the treatments on grass accuracy by workers who claimed the supervisors phone calls were helpful and those who claimed the calls were confusing during our end-line survey.[33]

The results of this analysis are presented in Table 6. Columns 1 and 2 are restricted to workers in the check-in treatment, and columns 3 and 4 are restricted to workers in the monitoring treatment. The analysis presented in all four columns include worker fixed effects. How useful workers in the check-in treatment thought their supervisor's regular calls were does not appear to influence the impact of the treatment calls on grass accuracy. This is somewhat reassuring given that the supervisor did not provide workers in the check-in treatment any feedback on how accurately they reported grass. In contrast, columns 3 and 4 demonstrate that the monitoring treatment had a positive effect on the accuracy of grass reporting for workers who had said the supervisor's calls improved their ability to do their job and a negative, but insignificant effect, on those who said the reverse. Perhaps not surprisingly, this implies that in order for monitoring to improve performance, it needs to be performed well. This finding is also broadly consistent with Nagin et al. (2002) by demonstrating that workers respond to monitoring interventions when they have positive attitudes towards the monitor.[34] More importantly for our purposes, these findings provide support for a general positive impact of increases in conspicuous monitoring, with the caveat that this increase might be best concentrated along task dimensions for which monitor evaluation errors are unlikely.

4.2. Differential attrition

In this section, we test whether our findings are robust to accounting for the differential attrition across groups. While in the previous discussion, we considered this differential attrition to be a relevant outcome of our treatment, it may also be driving our findings on other outcomes. Table A.2 compares the characteristics of attriters to those who remained on

[32] To generate our calculation, we use the cost of paying an M-Turk worker to evaluate a submission (0.05USD) and the supervisor's cost of time for each monitoring call (about 0.50USD). The supervisor in our setting was paid a salary and not paid per task or hour worked, but we estimate that each call took approximately 5 minutes and the supervisor's salary is based on an hourly rate of 10USD. Each call reported on all the previous day's submissions, so with a sample average of 6.42 submissions per day during the treatment period the cost of each monitoring call was approximately 0.821USD which we round up to 0.84USD to account for the cost of posting the task on MTurk. To avoid making strong assumptions about the potential welfare value of the information being generated by the workers in our setting, we use the cost of paying for a submission multiplied by the increased likelihood of it being a high quality submission in the monitoring group relative to the control as a measure of the returns from the monitoring intervention. On average, workers were paid 0.23USD per submission and, over the course of a week during treatment (the length of time between calls) monitored workers submitted about 4.5 fewer poor quality photos than control workers (using the raw data comparisons). Thus, each call saved the employer approximately 1.04USD.

[33] We do not see similar patterns in our other outcome variables, however, grass classifications were the only non-quantity based dimension on which the monitor provided precise feedback on. For the other quality dimensions mentioned in the monitoring calls, i.e. measures of picture quality, the monitor gave vaguer feedback. Moreover, while we did not anticipate that measuring grass classification accuracy would be more difficult than measuring photo quality, that does turn out to be the case. These two factors combined may explain why worker confusion over the calls seems to be most related to their grass classifications.

[34] We find similar results if we use a general measure of worker satisfaction with the job.

Table 6
Effect of monitoring activity on grass reporting accuracy by worker satisfaction with manager.

	Check-in group		Monitoring group	
	Negative (1)	Positive (2)	Negative (3)	Positive (4)
Treatment period	0.035	0.026	−0.042	0.095*
	(0.047)	(0.040)	(0.094)	(0.048)
Observations	1,112	2,304	2,633	3,589
R-squared	0.013	0.061	0.094	0.054
Mean dep var	0.626	0.604	0.606	0.669

Notes: This table reports estimated effects of treatment on the accuracy of grass reporting by treatment group type, and worker self-reported satisfaction with their supervisor's calls as part of the treatments. The sample of workers with negative worker experience are those who indicated their supervisor's regular phone calls made their job more confusing or made them invest less effort in the task. The sample of workers with positive worker experience are those who indicated their supervisor's regular phone calls made their job easier to do. All regressions include worker fixed effects. Standard errors clustered by worker reported in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

the job until the end of the study and demonstrates that attriters look very similar to non-attriters along all our measures of worker characteristics. Moreover, in the pre-treatment period, they perform similarly on the task.[35] This suggests that attrition was not motivated by worker ability or intended effort. A regression that estimates the effects of our treatments and worker characteristics on attriting further demonstrates that worker characteristics do not impact this outcome, but being in the monitoring treatment group does (Table A.3).

To investigate whether attrition is impacting our main findings, we re-run our main analysis on the sample of workers who stayed in the job until the end of period (Table A.4) and find that our interpretation of treatment effects change very little without this sample included, though the change in magnitudes in column 2, while not significant, does suggest that a potentially important mechanism through which monitoring generates higher overall outcomes is by increasing job retention. Moreover, the coefficient magnitude on the effect of the monitoring treatment in column 1 demonstrates that differences in quantity of submissions across groups during the treatment period is driven by differences in worker retention.

4.3. Mechanisms

In this section, we consider whether learning, expectations about rewards outside the current job, or changes in job satisfaction can explain why the monitoring treatment led to an increase in performance.

4.3.1. Learning

One possible reason that the monitoring treatment increases the quality of performance is because it provides learning benefits to these workers. As we note in Section 2, workers did receive intensive training on how to use the submission application, how to take useful pictures, and how to complete surveys accurately and by the end of the survey, we were convinced that all workers knew how to perform the task and understood what a high quality submission was across all dimensions. However, workers may have very quickly forgotten these lessons and the feedback on submission dimensions provided by the treatment calls could have helped workers remember how to perform their task or what task dimensions the employer valued.

There are several reasons why learning is unlikely to be driving our treatment effects. First, a learning explanation is difficult to reconcile with the increase in performance on task dimensions not discussed during the calls. Practically, it is not obvious why a reduction in photo blurriness or bad angles (monitored task dimensions) would generate a mechanical reduction in backlighting (a non-monitored task dimension) given that these two dimensions of photography are caused by distinct practices. Moreover, while the correlation between these two photo quality measures is positive and significant, even if we control for whether or not a photo is blurry or taken at a bad angle, the effect of monitoring on the non-monitored dimensions of photo quality is still significantly negative.

Second, given the simplicity of the job and the frequency of repetition, learning benefits should not disappear once they have been realized.[36] In contrast to what we would expect if learning explained our treatment effects, Fig. 3 demonstrates that the benefits of the monitoring treatment in terms of higher quality submissions fade over time. Table A.5 further demonstrates that the impact of the monitoring treatment on monitored photo quality dimensions is higher during than after treatment.[37]

[35] Attriters make no night submissions and non-attriters make statistically more, but the difference is economically very small.

[36] For instance, existing evidence demonstrates that people are unlikely to forget how to perform visual search or discrimination tasks that are frequently repeated once they learn how to perform them (Czerwinski et al., 1992; Stojanoski et al., 2018).

[37] Consistent with what we report in Table 6, the treatment effect of monitoring on grass accuracy is positive and significant during and positive though insignificant after treatment for workers who claimed the monitoring calls helped make their job easier, and negative, though insignificant, in both periods for those workers who claimed the opposite. Table A.6 reports these estimates.

We also test whether responses to our endline survey question about how difficult workers found explaining how to use the rewards and survey submission applications to other herdsman differ by treatment and control group. This question is useful for assessing whether workers in the monitoring group better understood how to perform the task than those in the other groups if we assume that a better understanding makes explaining the task to others easier (as demonstrated by Metzler and Woessmann, 2012). However, this analysis should be interpreted as suggestive because the endline survey was only completed by about 70% of workers and the workers that completed it were more engaged in the job than those who did not.[38] We find no evidence that those in the monitoring group had an easier time explaining how to submit surveys or understand what they would be paid for each submission than those in the check-in or control groups (see Table A.7). In summary, our data are not consistent with the monitoring treatment leading to performance increases because of worker learning.

4.3.2. Reciprocity

An alternative explanation for why the monitoring improved performance that is also consistent with a decline in treatment effects following the conclusion of the intervention is that the calls may have increased reciprocity considerations among workers. Monitored workers may have increased performance because of a change in their beliefs about the benefits to helping their supervisors unrelated to the current job's rewards. For instance, they may have interpreted the supervisor calls as a signal that the supervisor was evaluating or preparing them for future opportunities.

This explanation is inconsistent with a lack of increase in the quantity of submissions during treatment relative to pre-treatment in response to the calls. In particular, if workers were responding to non-job related considerations, the supervisor's signal about the quantity of submissions would likely have generated more submissions in both the check-in and monitoring groups despite there being no current job return from doing so. Similarly, we might expect that workers in the check-in treatment would have also responded by increasing the quantity of submissions in which they indicated the presence of grass.[39] Though, again, the number of submissions with grass was not a useful measure of performance from the employer's perspective, the calls may have led workers to believe the supervisor wanted the workers to focus more on submissions that indicate the presence of grass for reasons unrelated to of the current job incentive system. As we demonstrate in Appendix Table A.9, this is not occurring. Neither the check-in nor the monitoring treatment changed the number of submissions reported as having grass. Moreover, that we find no change in the accuracy of grass reporting in response to the check-in treatment (Table 5) demonstrates workers are not increasingly misreporting the presence of grass either.

To further test whether workers are responding to non-job related incentives to improve performance, we test whether monitored workers who state they would like to be re-hired by the employer for subsequent work in the endline survey increase their performance by more than those who do not want to be re-hired.[40] If workers believe their supervisor will be able to reward them with a subsequent hire, they may decide to improve their performance to increase their standing with the supervisor. We do not find consistent evidence that performance gains are concentrated among those monitored workers who want to continue on with the employer should an opportunity arise (see Table A.8).[41] In particular, while monitored workers who want to continue with the employer have a significantly larger treatment effect on monitored dimensions of quality, the treatment effect on non-monitored dimensions of quality is statistically the same. Thus, monitored workers improving their performance because of an increase in expectations about rewards outside of the current job does not provide a complete explanation for our treatment effects.

4.3.3. Job satisfaction

It is also possible that the monitoring treatment increased performance by improving how much workers enjoyed or valued the job. This improvement in job satisfaction could occur because workers who receive more regular feedback on their performance believe their supervisor is investing in monitoring them because their work is important[42] and, as a result, workers feel more valued (as suggested by Kim and Hamner, 1976; Rosen et al., 2006; Tziner and Latham, 1989). In contrast, the check-in treatment did not involve any performance feedback. Importantly, increased job satisfaction has been associated with higher levels of worker effort and performance (e.g. Christen et al., 2006).

This explanation is consistent with the lower level of worker attrition among the monitored group, and with an improvement in photo quality dimensions not mentioned during the monitoring call. It is also consistent with the gradual decline

[38] In particular, those who quit before the end of the task period are not included in the endline, and we were unable to survey all of those who did not quit. We do not have significant differential attrition between treatment and control groups (of those who did not quit, 60% of control, 79% of check-in, and 71% of monitoring finished survey), However, those who use their phones more often, and those with more herding experience are more likely to have completed the survey. In addition, we find that those who completed the endline survey submitted more completed photo surveys on average than those who did not across all three groups.

[39] These were the two task dimensions discussed in the check-in calls.

[40] We believe that our endline survey sample is particularly useful for analyzing this question because it is likely that the sample of workers who did respond to the survey are those most likely to have a desire to continue with the job in the future and that, thus, the mean response to this question is likely an upper bound on a desire to continue in the job.

[41] These estimates are similar if we include both workers who quit before the end of the task and those who did not fill in the endline in the "Do not want to continue" column.

[42] Similarly, monitored workers may believe their supervisor is higher ability than non-monitored workers do, which they may interpret as a signal that the employer valued their work enough to invest in a high quality supervisor.

in performance gains following the termination of the monitoring calls as workers began to suspect their supervisor was no longer monitoring their performance. Our finding that workers who had a poor experience with the monitoring calls experience lower performance gains than those with positive experiences further suggests that increased job satisfaction is an important mechanism through which increased visible monitoring activity can improve worker performance.

Further evidence in support of an increase in job satisfaction among monitoring workers is presented in the last two rows of Table A.7. In particular, we see some evidence that those in the monitoring group have are more likely to have a desire to continue with the job than those in the other groups. If we assume those who quit or did not complete the endline do not have a desire to continue, monitored workers are significantly more likely to want to be re-engaged for future work with the employer than control workers are.

Higher levels of job satisfaction should increase the cost of being terminated because the value of the job to the worker is higher. Thus, performance may have also increased with job satisfaction because of an increased fear of firing. Fear of firing may have increased independently of changes in job satisfaction because of change in monitored workers' perceptions about the intention of their supervisor to enforce the contract and, thus, fire them. While we cannot rule out fear of firing as an important mechanism for increased monitoring to improve worker performance, existing evidence suggests that increased job satisfaction is not consistent with an increased fear of firing (e.g. Clark, 2001; Origo and Pagani, 2009).

5. Conclusion

With changes in technology and globalization, alternative labor contracts that include work-from-home or remote work arrangements are becoming increasingly common (Bloom et al., 2015). These arrangements introduce novel managerial challenges that are not yet well understood (Bonet and Salvador, 2017). One of these challenges is how remote workers should be incentivized when their output is difficult to measure, for instance in multi-dimensional tasks where quantity and quality are important. This paper tests one possible solution to overcoming short-comings associated with performance-based pay for remote workers; increasing the salience of monitor activity on task dimensions not directly paid for.

To test the effects of our intervention, we run a field experiment among workers hired to collect, classify, and transmit data on rangeland conditions in Northern Kenya. Findings from our experiment demonstrate that workers who were randomly assigned to receive additional monitoring signals from their local manager increased performance on quality dimensions of the task discussed during these phones calls. Moreover, their performance on quality dimensions of the task not discussed during the calls also improved. We do not find that workers who were randomly assigned to receive additional communication from their local manager without changes in monitoring signals made economically significant changes to their performance. We also find that some of the treatment effects of enhanced monitoring persist after the treatment had ended, though in general they decline.

There are several boundary conditions on the external validity of our findings that we think are important to restate. First, given the physical limitations of signaling monitoring activity in the context of remote work, our monitoring intervention includes both a signal that workers' manager is observing dimensions of output not directly paid for and explicit feedback on dimensions of performance the employer cares about. Therefore, our findings are generalizable to settings where both observation and evaluation are being communicated to workers directly and with regular frequency. Second, given our findings on who improves the accuracy of their grass classifications, the impacts of monitoring are more positive when workers have a positive experience communicating with their supervisor than when they do not demonstrating that objective monitor quality is important for the success of this intervention. Third, while the population of workers in our setting have a clear understanding of market-based incentives[43], many have not been employed in full-time work because they are self-employed, and, thus, may be less accustomed to managerial oversight than remote workers in more traditional work settings. It is unclear whether this lack of experience would lead them to have relatively under- or over-estimated expectations about the frequency of supervisor monitoring, but it does suggest that the size of our estimated treatment effect may not precisely generalize to more traditional settings.

Our results are consistent with increased monitoring visibility without accompanied changes in incentive contracts increasing worker performance. Importantly, increasing communication frequency with managers without providing credible signals of monitoring does not achieve this effect. Combined, our analysis is consistent with the increased visibility of monitoring activity improving performance at least in part by increasing job satisfaction.

The results of our study have important implications for both the management of remote workers and for the use of worker performance data more generally. In particular, they demonstrate that in settings where workers have incomplete information about how they will be monitored, relatively low cost measures to increase the visibility of remote worker monitoring can lead to economically large changes in performance even without any changes to payment schemes or monitoring technologies. Our results also suggest that when tasks have output dimensions that cannot be affordably evaluated, managers may be able to improve worker performance even on these un-measurable dimensions by providing sufficiently high quality and frequent updates on the dimensions they are able to measure. Importantly, however, our results also demonstrate that this feedback will not improve performance if it only includes updates on task dimensions that are already being

[43] This is demonstrated in the quantity of their submissions, and is consistent with evidence from similar populations (Kelley et al., 2018; Lyons, 2020; Takahashi et al., 2016).

paid for directly or that are unrelated to the quality of performance. Furthermore, they suggest that selective use of fine-grained performance data in occasional performance reviews balances the need to ensure workers believe their performance is valued and accounted for with the importance of preserving worker autonomy and avoiding unnecessary disruptions (Bernstein, 2017), and that some active monitoring can, in fact, increase job satisfaction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Additional tables

Table A.1
Test for joint significance of worker characteristics on treatment assignment.

	(1) Check-in Treatment assignment	(2) Monitoring Treatment assignment
Age	0.002	0.029
	(0.021)	(0.034)
Current student	0.069	−0.107
	(0.189)	(0.448)
Highest level of education	−0.106	−0.157
	(0.074)	(0.128)
Years of herding experience	−0.005	−0.013
	(0.013)	(0.025)
Average number of calls per day	−0.001	−0.022
	(0.007)	(0.016)
Average number of SMS' sent per day	0.000	0.001
	(0.001)	(0.002)
F-test for joint orthogonality (*p*-Value)	0.874	0.509
Observations	77	77
R-squared	0.033	0.071
Mean dep var	1.442	1.883

Notes: This table reports the relationship between worker charateristics and treatment assignment. Worker Gender is excluded because all workers are Male. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.2
Worker characteristics and pre-treatment performance by early worker departure.

Worker characteristics	Remained until End of period	Left job before End of period	p-value^{+}
Panel A: Worker characteristics			
Age	22.39	22.182	0.853
	(0.359)	(0.872)	
Current Student	0.100	0.091	0.925
	(0.030)	(0.091)	
Highest level of education	2.070	2.273	0.448
	(0.082)	(0.304)	
Years of herding experience	13.280	13.727	0.801
	(0.547)	(1.907)	
Average number of calls per day	10.030	8.909	0.659
	(0.822)	(1.411)	
Average number of SMS' sent per day	42.060	30.091	0.526
	(5.934)	(17.278)	
Panel B: Output characteristics by worker			
Average number of submissions per day	7.814	7.730	0.726
	(0.075)	(0.190)	
High quality submissions per day	3.668	3.585	0.664
	(0.074)	(0.225)	
Proportion of days with at least ten submissions	0.528	0.537	0.746
	(0.011)	(0.033)	
Grass accuracy	0.583	0.567	0.367
	(0.006)	(90.017)	
Night time photo	0.002	0.000	0.007***
	(0.000)	(0)	
Poor quality monitored	0.123	0.142	0.093*
	(0.004)	(0.011)	
Poor quality not monitored	0.200	0.196	0.720
	(0.004)	(0.013)	

Notes: Standard deviations are in parentheses. Sample restricted to pre-treatment period. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3
Worker characteristics, treatment and early job exit.

DV: Left job before end of period	
Check-in	0.068
	(0.083)
Monitoring	−0.115**
	(0.052)
Age	−0.001
	(0.007)
Student	−0.027
	(0.108)
Education	0.031
	(0.041)
Herding experience	−0.000
	(0.006)
Average calls made per day	−0.003
	(0.003)
Average SMS sent per day	−0.000
	(0.001)
Observations	111
R-squared	0.072
Mean dep var	0.0991

Notes: This table reports estimated relationship between treatment and worker characteristics on leaving the job before the end of the study period. Robust standard errors reported in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4
Effect of monitoring activity on worker performance, restricted to non-attriters.

	(1) Number of submissions	(2) High quality photos per day	(3) Grass accuracy	(4) Night submissions	(5) Bad quality monitored	(6) Bad quality not Monitored
Treatment period	−0.968**	0.020	0.061**	0.003	0.014	0.027**
	(0.401)	(0.275)	(0.030)	(0.004)	(0.012)	(0.011)
Check-in ×	0.596	0.552	−0.021	0.003	−0.029*	−0.015
Treatment period	(0.785)	(0.488)	(0.039)	(0.005)	(0.016)	(0.014)
Monitoring ×	0.039	0.216	−0.034	0.001	−0.078***	−0.040**
Treatment period	(0.691)	(0.303)	(0.496)	(0.004)	(0.017)	(0.018)
Check-in × Treatment=						
Monitoring × Treatment	0.978	0.562	0.794	0.504	0.002	0.144
Observations	11,995	10,951	23,946	86,994	31,306	31,306
R-squared	0.364	0.191	0.059	0.023	0.037	0.008
Mean dep var	7.253	3.825	0.615	0.00554	0.106	0.205

Notes: This table reports estimated effects of treatment on dimensions of the task that were and were not mentioned in the monitoring treatment phone calls during and after treatment was occurring. The sample is restricted to workers who did not leave the study before the end of the period. In columns 1 and 2, an observation is a worker day and regressions include worker fixed effects. In columns 4–7, an observation is a submission. Robust standard errors are reported in parentheses. High quality submissions per day is the number of submissions not classified as bad quality along any dimension of quality and is restricted to submissions reviewed by external reviewers and weighted by the proportion of photos per participant day that were reviewed by external reviewers. Grass accuracy is equal to one if a submission accurately reports the presence of grass and zero otherwise. Night submission is equal to one if the photo included in the submission was taken at night and zero otherwise. Worker departure is equal to one if a worker quit or disappeared from the job before the end of the study period and zero otherwise. Columns 3 and 4 include worker fixed effects and columns 5 and 6 include site fixed effects. Standard errors clustered by worker in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5
Effect of monitoring activity on worker performance during and after treatment.

Panel A: Quantity outcomes		
	Number of submissions (1)	High quality photos per day (2)
During treatment	−1.617***	−0.317
	(0.463)	(0.288)
Post treatment	−3.445***	−1.563***
	(0.475)	(0.277)
Check-in*	0.173	0.233
During treatment	(0.869)	(0.519)
Check-in*	1.859*	0.746
Post-treatment	(0.975)	(0.536)
Monitoring*	0.726	0.581
During treatment	(0.729)	(0.501)
Monitoring*	1.836**	1.165**
Post-treatment	(0.857)	(0.538)
Monitoring-in × During treatment=		
Monitoring × Post-treatment	0.074*	0.118
Observations	16,091	16,091
R-squared	0.400	0.363
Mean dep var	6.667	0.282

(*continued on next page*)

Table A.5 (*continued*)

Panel B: Quality dimensions	Grass accuracy (1)	Night submissions (2)	Bad quality monitored (3)	Bad quality not monitored (4)
During treatment	0.064**	0.003	0.010	0.028***
	(0.029)	(0.004)	(0.013)	(0.010)
Post treatment	0.041	0.004	0.013	0.024
	(0.035)	(0.007)	(0.026)	(0.019)
Check-in*	−0.026	0.002	−0.024	−0.013
During treatment	(0.039)	(0.005)	(0.016)	(0.014)
Check-in*	−0.063	0.004	0.014	0.007
Post-treatment	(0.056)	(0.008)	(0.029)	(0.021)
Monitoring*	−0.041	0.003	−0.073***	−0.040**
During treatment	(0.051)	(0.005)	(0.017)	(0.017)
Monitoring*	−0.030	0.013	−0.037	−0.067***
Post-treatment	(0.054)	(0.013)	(0.036)	(0.024)
Monitoring-In × During treatment=				
Monitoring × Post-treatment	0.770	0.350	0.218	0.219
Observations	28,570	107,286	38,010	38,010
R-squared	0.057	0.036	0.039	0.009
Mean dep var	0.611	0.00668	0.110	0.204

Notes: This table reports estimated effects of treatment on dimensions of the task that were and were not mentioned in the monitoring treatment phone calls during and after treatment was occurring. In Panel A, an observation is a worker day. In Panel B, an observation is a submission. Columns 4 and 5 of Panel C are at the worker level. Robust standard errors are reported in parentheses. High Quality Submissions per Day is the number of submissions not classified as bad quality along any dimension of quality and is restricted to submissions reviewed by external reviewers and weighted by the proportion of photos per participant day that were reviewed by external reviewers. Grass accuracy is equal to one if a submission accurately reports the presence of grass and zero otherwise. Night submission is equal to one if the photo included in the submission was taken at night and zero otherwise. Worker Departure is equal to one if a worker quit or disappeared from the job before the end of the study period and zero otherwise. All columns in Panel A include worker fixed effects. Columns 1 and 2 in Panel B include worker fixed effects and columns 3 and 4 in Panel B include site fixed effects. Standard Errors clustered by worker are in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6
Effect of monitoring activity on grass reporting accuracy persistence by worker satisfaction with manager.

	(1) Check-in group	(2)	(3) Monitoring group	(4)
	Negative	Positive	Negative	Positive
During treatment	0.034	0.035	−0.039	0.092*
	(0.046)	(0.043)	(0.094)	(0.048)
Post treatment	−0.031	−0.003	−0.045	0.074
	(0.050)	(0.074)	(0.079)	(0.064)
Observations	1439	3176	2926	4165
R-squared	0.020	0.065	0.093	0.051
Mean dep var	0.603	0.598	0.598	0.673

Notes: This table reports estimated effects of treatment on the accuracy of grass reporting during and after treatment by treatment group type, and worker self-reported satisfaction with their supervisor's calls as part of the treatments. The sample of workers with negative worker experience are those who indicated their supervisor's regular phone calls made their job more confusing or made them invest less effort in the task. The sample of workers with positive worker experience are those who indicated their supervisor's regular phone calls made their job easier to do. All regressions include individual fixed effects. Robust standard errors reported in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.7
Endline survey responses by treatment and control groups.

	Control	Check-in	Monitoring	Significant differences?
Explaining rewards and survey submission was hard or very hard	0.167 (0.078)	0.091 (0.063)	0.167 (0.078)	None
Desire to continue in any future tasks	0.708 (0.095)	0.636 (0.105)	0.833 (0.078)	None
Observations	24	22	24	
Desire to continue in any future tasks (assume "no" if survey not completed)	0.378 (0.073)	0.412 (0.086)	0.589 (0.086)	Monitoring-control > 0
Observations	45	34	34	

Notes: The full survey questions analyzed in this table are reported in the appendix of the main paper. The *Significant differences?* reports whether *t*-tests of differences between any of the three groups are significant.

Table A.8
Quality of submissions by those who want a subsequent job opportunity from employer.

	(1)	(2)	(3)	(4)
	Blurry or bad angle		Backlit or bad quality other	
	No continue	Continue	No continue	Continue
Treatment period	−0.037*	0.056***	−0.001	0.046***
	(0.021)	(0.017)	(0.017)	(0.015)
Managerial activity ×	0.022	−0.058**	−0.022	−0.024
Treatment period	(0.027)	(0.022)	(0.021)	(0.019)
Monitoring ×	−0.022	−0.112***	−0.073**	−0.061***
Treatment period	(0.023)	(0.025)	(0.032)	(0.021)
Monitoring × Treatment coefficient equality across regressions	0.001**		0.734	
Observations	6323	15,688	6323	15,688
R-squared	0.012	0.047	0.010	0.010
Mean dep var	0.0805	0.110	0.201	0.204

Notes: This table reports estimated effects of treatment on monitored and non-monitored dimensions of quality. Treatment is equal to 1 during the period that the monitoring or check-in treatments are occurring. Monitoring is equal to one if a worker received monitoring phone calls from his supervisor. Check-in is equal to one if a worker received check-in phone calls from his supervisor. An observation is a worker day. Columns 1 and 3 restrict the sample to workers who report not wanting to continue with the employer if a future opportunity arises in the endline survey. Columns 2 and 4 restrict the sample to workers who report wanting to continue with the employer if a future opportunity arises in the endline survey. All regressions include worker fixed effects. Standard Errors clustered by worker are in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%%.

Table A.9
Grass reporting by treatment groups.

DV: Whether or not submission reported presence of grass	
Treatment period	2.117***
	(0.369)
Check-in × Treatment period	0.003
	(0.497)
Monitoring × Treatment period	−0.186
	(0.504)
Total submissions	0.505***
	(0.031)
Observations	13,311
R-squared	0.638
Mean dep var	3.428

Notes: This table reports estimated effects of treatment on the number of submissions reported to have grass per day. Treatment is equal to 1 during the period that the monitoring or check-in treatments are occurring. Monitoring is equal to one if a worker received monitoring phone calls from his supervisor. Check-in is equal to one if a worker received check-in phone calls from his supervisor. An observation is a worker day. Worker fixed effects are included. Standard Errors clustered by worker are in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

References

Al-Ubaydli, O., Andersen, S., Gneezy, U., List, J.A., 2015. Carrots that look like sticks: Toward an understanding of multitasking incentive schemes. South. Econ. J. 81 (3), 538–561.

Alchian, A.A., Demsetz, H., 1972. Production, information costs, and economic organization. Am. Econ. Rev. 62 (5), 777–795.

Anteby, M., Chan, C.K., 2018. A self-fulfilling cycle of coercive surveillance: Workers invisibility practices and managerial justification. Organ. Sci. 29 (2), 247–263.

Baker, G., Gibbons, R., Murphy, K.J., 1994. Subjective performance measures in optimal incentive contracts. Q. J. Econ. 109 (4), 1125–1156.

Baker, G.P., 1992. Incentive contracts and performance measurement. J. Polit. Econ. 100 (3), 598–614.

Bernstein, E.S., 2012. The transparency paradox: A role for privacy in organizational learning and operational control. Adm. Sci. Q. 57 (2), 181–216.

Bernstein, E.S., 2017. Making transparency transparent: The evolution of observation in management theory. Acad. Manag. Ann. 11 (1), 217–266.

Bilal, N.K., Herbst, C.H., Zhao, F., Soucat, A., Lemiere, C., 2011. Health extension workers in ethiopia: Improved access and coverage for the rural poor. Yes Africa Can: Success Stories from a Dynamic Continent. The World Bank, pp. 433–443.

Bloom, N., Liang, J., Roberts, J., Ying, Z.J., 2015. Does working from home work? Evidence from a chinese experiment. Q. J. Econ. 165, 218.

Bloom, N., Sadun, R., Van Reenen, J., 2016. Management as a Technology? No. w22327. National Bureau of Economic Research. (16-133)

Bonet, R., Salvador, F., 2017. When the boss is away: Manager–worker separation and worker performance in a multisite software maintenance organization. Organ. Sci. 28 (2), 244–261.

Bresnahan, T.F., Brynjolfsson, E., Hitt, L.M., 2002. Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. Q. J. Econ. 117 (1), 339–376.

Bruhn, M., McKenzie, D., 2009. In pursuit of balance: Randomization in practice in development field experiments. Am. Econ. J. Appl. Econ. 1 (4), 200–232.

Christen, M., Iyer, G., Soberman, D., 2006. Job satisfaction, job performance, and effort: A reexamination using agency theory. J. Market. 70 (1), 137–150.

Clark, A.E., 2001. What really matters in a job? Hedonic measurement using quit data. Labour Econ. 8 (2), 223–242.

Courty, P., Marschke, G., 2004. An empirical investigation of gaming responses to explicit performance incentives. J. Labor Econ. 22 (1), 23–56.

Cragg, M., 1997. Performance incentives in the public sector: Evidence from the job training partnership act. J. Law Econ. Organ. 13 (1), 147–168.

Czerwinski, M., Lightfoot, N., Shiffrin, R.M., 1992. Automatization and training in visual search. Am. J. Psychol. 105 (2), 271–315.

De Jong, B.A., Dirks, K.T., 2012. Beyond shared perceptions of trust and monitoring in teams: Implications of asymmetry and dissensus.. J. Appl. Psychol. 97 (2), 391.

de Rochambeau, G., 2017. Monitoring and intrinsic motivation. IGC Working Paper.

Dihel, N., 2011. Beyond the Nakumatt Generation: Distribution Services in East Africa. The World Bank: Africa Trade Policy Notes (26). The World Bank.

Lu, S.F., 2012. Multitasking, information disclosure, and product quality: Evidence from nursing homes. J. Econ. Manag. Strat. 21 (3), 673–705.

Frey, B.S., 1993. Does monitoring increase work effort? The rivalry with trust and loyalty. Econ. Inq. 31 (4), 663–670.

Gibbons, R., 1998. Incentives in organizations. J. Econ. Perspect. 12 (4), 115–132.

Holmstrom, B., Milgrom, P., 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. J. Law Econ. Organ. 7, 24–52.

Hong, F., Hossain, T., List, J.A., Tanaka, M., 2013. Testing the Theory of Multitasking: Evidence From a Natural Field Experiment in Chinese Factories. Technical Report. National Bureau of Economic Research.

Jarvenpaa, S.L., Leidner, D.E., 1999. Communication and trust in global virtual teams. Organ. Sci. 10 (6), 791–815.

Katz, L.F., 1986. Efficiency wage theories: A partial evaluation. NBER Macroecon. Ann. 1, 235–276.

Kelley, E., Lane, G., Schönholzer, D., 2018. The Impact of Monitoring Technologies on Contracts and Employee Behavior: Experimental Evidence from Kenyas Transit Industry. Technical Report. Mimeo, Berkeley.

Kim, J.S., Hamner, W.C., 1976. Effect of performance feedback and goal setting on productivity and satisfaction in an organizational setting.. J. Appl. Psychol. 61 (1), 48.

Larkin, I., Pierce, L., 2015. Compensation and employee misconduct: The inseparability of productive and counterproductive behavior in firms. In: Organizational Wrongdoing: Key Perspectives and New Directions, pp. 1–27.

Lazear, E.P., 1986. Salaries and piece rates. J. Bus. 59 (3), 405–431.

Lazear, E.P., 1991. Labor economics and the psychology of organizations. J. Econ. Perspect. 5 (2), 89–110.

Lyons, E., 2020. The impact of job training on temporary worker performance: Field experimental evidence from insurance sales agents. J. Econ. Manag. Strat. 29 (1), 122–146.

Maskin, E., Tirole, J., 1990. The principal-agent relationship with an informed principal: The case of private values. Econom. J. Econom. Soc. 379–409.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D., 2012. Big data: the management revolution. Harvard Bus. Rev. 90 (10), 60–68.

McPeak, J.G., Barrett, C.B., 2001. Differential risk exposure and stochastic poverty traps among east african pastoralists. Am. J. Agric. Econ. 83 (3), 674–679.

Metzler, J., Woessmann, L., 2012. The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. J. Dev. Econ. 99 (2), 486–496.

Nagin, D.S., Rebitzer, J.B., Sanders, S., Taylor, L.J., 2002. Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. Am. Econ. Rev. 92 (4), 850–873.

Neuwirth, B., 2014. Marketing Channel Strategies in Rural Emerging Markets. Kellogg School of Management Working Paper 22. Kellogg School of Management, Northwestern University.

Origo, F., Pagani, L., 2009. Flexicurity and job satisfaction in Europe: The importance of perceived and actual job stability for well-being at work. Labour Econ. 16 (5), 547–555.

Pierce, L., Snow, D.C., McAfee, A., 2015. Cleaning house: The impact of information technology monitoring on employee theft and productivity. Manag. Sci. 61 (10), 2299–2319.

Prendergast, C., 2002. The tenuous trade-off between risk and incentives. J. Polit. Econ. 110 (5), 1071–1102.

Ranganathan, A., Benson, A., 2016. Hemming and Hawing Over Hawthorne: Work Complexity and the Divergent effects of Monitoring on Productivity. Technical Report, Working paper. Stanford Graduate School of Business, Stanford, CA.

Reardon, T., Timmer, C.P., Barrett, C.B., Berdegué, J., 2003. The rise of supermarkets in Africa, Asia, and Latin America. Am. J. Agric. Econ. 85 (5), 1140–1146.

Rosen, C.C., Levy, P.E., Hall, R.J., 2006. Placing perceptions of politics in the context of the feedback environment, employee attitudes, and job performance.. J. Appl. Psychol. 91 (1), 211.

Savin, N.E., 1984. Multiple hypothesis testing. Handb. Econom. 2, 827–879.

Staats, B.R., Dai, H., Hofmann, D., Milkman, K.L., 2016. Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. Manag. Sci. 63 (5), 1563–1585.

Stojanoski, B., Lyons, K.M., Pearce, A.A., Owen, A.M., 2018. Targeted training: Converging evidence against the transferable benefits of online brain training on cognitive function. Neuropsychologia 117, 541–550.

Takahashi, K., Ikegami, M., Sheahan, M., Barrett, C.B., 2016. Experimental evidence on the drivers of index-based livestock insurance demand in southern ethiopia. World Dev. 78, 324–340.

The World Bank, 2019. Agriculture and Food: Overview. Technical Report. The World Bank.

Tziner, A., Latham, G.P., 1989. The effects of appraisal instrument, feedback and goal-setting on worker satisfaction and commitment. J. Organ. Behav. 10 (2), 145–153.

Warkentin, M.E., Sayeed, L., Hightower, R., 1997. Virtual teams versus face-to-face teams: An exploratory study of a web-based conference system. Decis. Sci. 28 (4), 975–996.

Weisbrod, B.A., 1989. Rewarding performance that is hard to measure: The private nonprofit sector. Science 5, 244.