Monitoring A, While Hoping for A & B: Experimental Evidence from a Multidimensional Task[*]

Preliminary and incomplete. Please do not distribute.

Nathan Jensen[†] Elizabeth Lyons[‡] Eddy Chebelyon[§] Ronan Le Bras[¶]

Carla Gomes[||]

December 19, 2016

Abstract

Monitoring workers in order to match rewards to performance is a central justification for the importance of organizations and of management. However, when both inputs and outputs are difficult to measure, achieving this objective is difficult. We consider whether signaling and demonstrating monitor productivity on some performance dimensions leads to an improvement in worker performance on all dimensions. Preliminary results from a field experiment run among remote multi-dimensional task workers in rural Kenya demonstrate that increasing the visibility of monitoring on some dimensions improves performance on most of those dimensions, as well as performance on others. Our evidence is consistent with this monitoring acting as a signal of managerial productivity.

JEL Classification: J24, M54, D83, O13

[*] 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, UCSD's Rady School of Management, and the 4th International Conference on Computational Sustainability for valuable input and feedback. Oscar Naibei provided valuable 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 author(s) and not necessarily those of the Commonwealth of Australia. The Commonwealth of Australia accepts no responsibility for any loss, damage or injury resulting from reliance on any of the information or views contained in this publication. 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"

[†] Dyson School of Applied Economics and Management, Cornell University, 438 Warren Hall, Ithaca, NY 14850. ndj6@cornell.edu

[‡] School of Global Policy & Strategy, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093. lizlyons@ucsd.edu.

[§] School of Global Policy & Strategy, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093. echebelyon@ucsd.edu

[¶] Department of Computer Science, Cornell University, 344 Bill and Melinda Gates Hall, Ithaca, NY 14830. lebras@cs.cornell.edu

[||] Department of Computer Science, Cornell University, 353 Bill and Melinda Gates Hall, Ithaca, NY 14830. cmw84@cornell.edu

1 Introduction

The importance and difficulty of monitoring workers in order to properly reward and punish them is a major justification for the existence of organizations (Alchian and Demsetz, 1972).[1] As remote work becomes increasingly common (e.g. Bloom et al., 2015), how monitoring occurs is evolving. For instance, firms are increasingly using software to monitor worker inputs and outputs (Bresnahan et al., 2002). However, there are task types that may be difficult to track through IT programs. In particular, monitoring inputs of remote work not performed online is practically challenging. As a result, even when optimal inputs are definable, input-based incentive pay may not be optimal because inputs cannot be verified (Prendergast, 2002). Holmstrom and Milgrom (1991) argue that employees who work from home should receive more performance-based pay than those who work in the organization's offices to overcome difficulties associated with monitoring inputs. However, if output cannot be effectively measured, performance-based incentives may not be optimal either.[2]

In this paper we test one possible solution to overcoming difficulties associated with managing workers' whose inputs and outputs are costly to measure. In particular, we test whether more active and visible monitoring of relatively easy to measure aspects of worker output improves worker performance in all areas. Theoretically, this could have two opposing affects on performance. On one hand, increasing the visibility and intensity of monitoring may act as a signal to workers that they are being actively tracked. If workers believe that active monitoring on some dimensions is positively correlated with active monitoring on other dimensions of performance, for instance

[1] For instance, 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.

[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).

because active monitors are more productive, this form of monitoring could increase overall performance (e.g. Spence, 1973). Thus, unlike an incentive scheme in which some task dimensions are rewarded more than others as in Holmstrom and Milgrom (1991), workers interpret monitoring on some dimensions as a signal of general management capabilities rather than as a signal of what dimensions they will be rewarded or punished for. On the other hand, if workers believe active monitoring on some dimensions is negatively correlated with active monitoring on other dimensions, for instance because managers value some dimensions more than others, this form of monitoring could increase the dimensions of performance being monitored, and decrease others (e.g. Kerr, 1975).

To test whether increasing the visibility of monitoring on some dimensions affects performance, we designed and implemented a field experiment on a population of workers who are particularly difficult to monitor. Specifically, we randomly assigned a monitoring treatment and a managerial activity treatment to workers tasked with collecting, classifying, and transmitting data on rangeland conditions in rural areas of Northern Kenya. Importantly, these workers submit large quantities of data based on information not available to managers. As a result, the quantity and characteristics of output are easy to monitor, but the quality is not.

In the both monitoring and managerial activity treatment groups, workers received a phone call from their local manager once every five days. In the managerial activity group, the local manager told workers how much data had been received the previous day, and how much of that data had a specific characteristic, in particular, how many of the data points were reported to have grass in them. 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 managerial activity group, and also told the worker how much of the received data was poor on two dimensions of quality. These two treatment groups allow us to differentiate between the effects of a call acting as a reminder or time waster and the effects of actual monitoring.[3] We observe workers over a 149 day period. In order to observe how the treatments affect an individual worker's performance and whether the effects continue after monitoring stops, we begin treating workers in the treatment groups 49 days into the sample period, and stop treating them before the end of the sample period.[4]

Preliminary findings from this experiment are consistent with monitoring on some dimensions increasing performance on those and other dimensions. In particular, we find that worker effort improves on most of the task dimensions they were being activity monitored on and that it improves on those they were not being activity monitored on as well. We find that these effects are significantly larger than when workers observe an increase in managerial actively without a change in active monitoring. Moreover, we find that these effects persist and become larger after treatment ends suggesting workers learn about how to perform from repeated activity.

Understanding how organizations can overcome high monitoring costs without sacrificing performance incentives is critical for organizational success, particularly as production modes are changing in response to globalization and advances in technologies. This project provides causal evidence on whether monitoring on easy to observe dimensions of output quality signals high manager productivity and creates incentives for workers to increase quality overall, or whether it

[3] For instance, given that the workers have virtually no contact with managers while working, the manager's call may remind workers that they have a manager or that their data is being transmitted to a server. If it is active monitoring that leads to a change in performance as opposed to this reminder, we should see the monitoring treatment have an effect and the managerial activity treatment have no effect.

[4] There are 4 different stop times to determine if length of treatment matters.

signals managers are only monitoring on those dimensions and crowds out incentives for workers to invest effort on other dimensions. Our findings provide empirical support for the former. Unlike evidence that providing performance-based incentives for one dimension of a task leads workers to switch from un-incentivized dimensions to incentivized ones (e.g. Holmstrom and Milgrom, 1991), monitoring on some dimensions appears to increase performance overall. This suggests that when observing and measuring inputs and outputs is not possible, increasing the visibility of manager productivity may overcome the deficiencies associated with performance-based pay in these situations.

In addition to contributing to the organizational and labor economics literature, 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 including monitoring difficulties (Bilal et al., 2011). Low-cost monitoring solutions for remote work may therefore have significant implications for private sector development in lower income countries.

This paper proceeds as follows. Section 2 reviews related literature to motivate the study; section 3 presents the experiment design; section 4 describes and summarizes the data; section 5 presents the empirical results from the experiment; and section 6 summarizes and concludes.

2 Literature on Monitoring & Incentives

Literature on the importance of motivating workers to perform their tasks as instructed is far too substantial and broad to review here. This section will focus on a narrow subset of this literature in Economics and Management that considers questions about how incentives can accurately reward performance, particularly for jobs that have multiple performance dimensions that matter to the manager and for those with hard-to-measure performance dimensions. In addition, literature on worker response to monitoring is reviewed.

A relatively early analysis of difficulties associated with providing incentives for workers in multidimensional jobs is presented in Kerr (1975) who provides several specific examples of incentive systems that reward one dimension of performance more strongly than another leading to poor performance on the less incentivized dimension. Baker et al. (1988) also note trade-offs between quality and quantity in piece-rate payment schemes and the potential for incentives tied to objective performance metrics to crowd out subjective dimensions of performance. Holmstrom and Milgrom (1991) formalize the difficulties associated with providing incentives for workers in multidimensional tasks in order to explain a number of observed features of employment including the pervasiveness of fixed wages. Importantly for our study, Holmstrom and Milgrom (1991) find that when re-allocating fixed effort across task dimensions increases effort on one dimension at the expense of effort exerted on others and when performance on some dimensions is difficult to measure, providing high powered incentives for the easy to measure dimensions of performance leads to poorer performance on the others. Several practical implications follow from this including that remote work should involve more performance-based pay than on-site work in order to reduce incentives to spend time on personal activities, and that when quality and quantity outcomes both matter and quality is hard to measure, incentives for quantity may not be optimal. This leads to a difficult trade-off for remote work incentives when the quality and quantity of output matters and quality is difficult to measure.

There is a growing empirical literature on incentives for multi-dimensional tasks that, for the most part, supports the theoretical evidence in Holmstrom and Milgrom (1991).[5] For instance, using evidence from a field experiment run in Chinese factories, Hong et al. (2013) find that providing incentives for increased quantity reduces the quality of output. Similarly, Dumont et al. (2008) use data on changes in physician activities in response to a decrease in output-based pay to show that under the output-based pay scheme, physicians performed more services and appear to have done so at the expense of quality of care and outside activities like teaching. Englmaier et al. (2012) also show a trade-off between quantity and quality of output when quantity-based pay is made more salient.[6]

Without focusing explicitly on multi-dimensional tasks, a number of theoretical papers have considered whether to provide input or output-based incentives depending on whether inputs or outputs are easier to measure.[7] For example, Lazear (1986) models trade-offs between the cost of monitoring output as required by output-based pay and the persistence of low output workers in input-based work and finds that input-based pay reduces difficulties and costs associated with monitoring output but also lowers sorting efficiency. Results from a model developed in Baker

[5] One exception, for example, is Mullen et al. (2010) who do not find evidence that pay-for-performance among physicians affected any dimensions of quality.

[6] Additional support for incentives leading to trade-offs across task dimensions can be found in Forbes et al. (2015), Rubin et al. (2016), and Slade (1996).

[7] Much of this work is described in more detail in Gibbons (1998).

(1992) demonstrate that output-based pay may dominate when managers can 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. Similarly, Prendergast (2002)'s theory shows that higher uncertainty about worker effort is associated with an increased likelihood of ouput-based payment schemes. Empirical evidence on the relationship between monitoring and payment schemes largely support these theories (e.g. Courty and Marschke, 2004; Cragg, 1997).

In general, the literature discussed here 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. However, it also suggests that when workers have information that managers do not and when worker inputs cannot be verified, output-based pay likely dominates. Therefore, in settings like the one studied in this paper where output is difficult to measure well, worker effort is uncertain, and there is asymmetric information between workers and managers, it is unclear how payment schemes should be structured. The alternative we study in this paper is to combine input-based and output-based incentives schemes by providing output-based pay on dimensions of output that are easily measured and attempting to increase fear of firing due to shirking on other dimensions by increasing managerial and monitoring activity.

Importantly, we are not the first to test how workers respond to monitoring activity and monitor productivity. Probably most related to our paper is Al-Ubaydli et al. (2015) which also tests 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 monitor is productive. Also closely related to our paper, Feng 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.[8] Additionally, literature on the trade-offs between trust and monitoring demonstrates that monitoring may reduce worker performance even on the monitored dimensions of performance if workers interpret it asx manager distrust (e.g. De Jong and Dirks, 2012; Frey, 1993). This suggests that our monitoring intervention could reduce performance on all dimensions if it crowds out overall effort. Our work contributes to this empirical literature on worker responses to monitoring by testing whether increasing monitoring on some performance dimensions is interpreted as a signal of the performance dimensions that matter to the monitor or as a signal of monitor productivity overall, and whether these signals impact performance. As in Al-Ubaydli et al. (2015), we test this in a setting with uncertainty about both monitor productivity and about worker effort.

3 Experimental Design

To test whether active and visible monitoring of easy-to-measure dimensions of output changes 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.

[8] This is distinct from our setting where we are altering monitoring activity on dimensions of performance that workers are not receiving direct incentives for.

3.1 Study Setting and Population

We ran our experiment on 113 workers hired to collect and transmit information on rangeland conditions in rural areas of Central Kenya over a 149 day period. Workers were located in five divisions; two in Samburu County, two in Isiolo County, and one in Laikipia County. Figure 1 demonstrates the region these workers covered. The 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 crowd-sourcing as a means for improving resource allocation among pastoralist communities.[9] 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.[10]

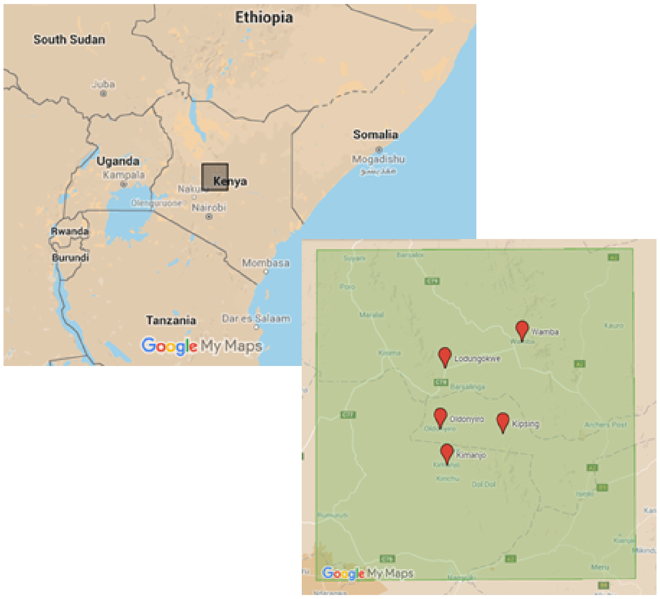

Figure 1: Region Covered by Study Participants

In order to collect and transmit information on rangeland conditions, pastoralists were sup-

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

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

plied with smartphones that included cameras and GPS. A crowd-sourcing mobile application was developed for the purpose of this job, and pastoralists were to submit all their data through the application. To achieve a single completed data point input, workers were required to take a photo in the application and then select whether the rangeland in the photo includes any grass, trees, or bushes, and, if so, whether each is green or brown in color. In addition, workers were required to indicate carrying capacity of the rangeland for cattle.[11] Workers could be paid between $0.05 and $0.40 per submission for up to ten photo and classification submissions per day (referred to as a survey submission going forward).[12] 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. Workers received three days of intensive training on the use of the smartphone, the application, and the task. They were employed on this job between March and August of 2015, and none of the workers were fired for any reason.

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 which classifications were made are automatically recorded with the data 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 photo are difficult to verify because of the large quantity

[11] Some of the 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 on the application that corresponded to their responses.

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

of data submitted. Workers may have an incentive to mis-classify photos to reduce the time it takes to submit each one, for instance because choosing the first option on each screen in the application would be faster than choosing the correct option, or to submit quickly taken, poor quality photos. In addition, if they believe that aid to the region would be affected by the crowdsourcing effort,[13] then they may have an incentive to classify photos as indicating rangeland conditions are worse than they are in reality.

3.2 Experimental Treatments

To test whether increasing the visibility and activity of monitoring on some task dimensions affected the performance of workers, we introduced two managerial treatments. Workers assigned the first treatment, which we will hereafter refer to as the "managerial activity" treatment, received a call from their manager every five days. During the call, the manager told each worker how many submissions they had made the previous day, and how many of those submissions were classified as having grass in them. The manager did not give workers any evaluation-based feedback on the quality or quantity of data received, and in particular, did not tell workers whether the photos were correctly specified as having grass in them. Workers assigned the second treatment, which we will hereafter refer to as the "monitoring" treatment, also received a call from their manager every five days. The beginning of the call was identical to the call in the managerial activity treatment. However, workers in this treatment group were also told which submissions from the prior day had correctly and incorrectly classified the presence of grass in the photo. In addition, the manager

[13] 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 distribution among other things.

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:

> *Managerial Activity 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. When we examined the photos. 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.

3.3 Study Implementation

The managerial activity and monitoring treatments were each 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 division to ensure that each division has workers in all three groups. Each day, the manager called all workers in the treatment groups in a single division resulting in one division being called 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 whether the stickiness of the treatments depends on how long the treatment period is, 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. The questionnaire asked about their educational and work backgrounds, their demographics, and their 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.

4 Data & Estimation Strategy

4.1 Data Description

Our study includes data from 149 days of worker activity from March to August in 2015 and our experiment ran from April 24 to July 26 of that year. Our sample includes the population of workers hired for the rangeland crowdsourcing project. Of our total population of 113 workers, 45 were assigned to the control group, 34 to the managerial activity group, and 34 to the monitoring group. 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 they 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. In addition, we 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. 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.[14]

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 [15], 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 four variables summarized in Panel B, specifically the proportion of dates with

[14] We are in the process of collecting data on the accuracy of photo classifications from a Mechanical Turk project for which Turkers are to classify photo submissions in the same way the workers did, and classify whether each photo was blurry, or of general bad quality. In an effort to reduce classification errors from the Mechanical Turk project, a subset of photos are being classified by at least two Turkers. At the time of writing, 21% of photos were checked by Turkers at least once.

[15] Including whatsapp, facebook and viber messages

no submissions, the average number of submissions per day, the proportion of days with at least ten submissions, and whether or not 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, on average workers rarely did not submit a photo per day and they submitted approximately the maximum number they would be paid for. The remaining variables summarized are at the submission level. We received just over 107,00 submissions, of which about 90% were classified as having shrubs, 50% were classified as having grass, and almost all were classified as having trees. Less than 1% of photos were taken outside of the permissible hours. Poor quality photo is equal to one if a submission reviewer indicated the photo was back lit or was poor quality for some other unspecified reason, and zero otherwise. At the time of writing, we have this information on about 23,000 submissions of which about 20% were of poor quality.

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 pairwise comparisons of the groups or a test of mean equality across the three groups, which we report the p-values from in Table 2. These tests confirm that workers in each group are comparable, and suggest that our random assignment of workers to treatments was appropriately executed.

We also compare average outcomes across treatment groups over the course of our study. In particular, we consider outcome characteristics that were mentioned in the monitoring treatment and those that were not. The features of submissions that were mentioned in the monitoring treatment were the number of submissions made in a day, the accuracy of grass reporting, the time the

	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					
Proportion of Dates with No Submissions	0.085	0.119	0.002	0.613	113
Average Number of Submissions per Day	9.400	2.314	2.959	14.579	113
Proportion of Days with At Least Ten Submissions	0.654	0.242	0.059	0.970	113
Worker Left Job	0.097	0.298	0	1	113
Grass Reported	0.501	0.500	0	1	107,286
Shrubs Reported	0.893	0.308	0	1	107,286
Trees Reported	0.972	0.165	0	1	107,286
Night Time Photo	0.007	0.081	0	1	107,286
Poor Quality Photo	0.200	0.400	0	1	23,013

Table 1: Worker Summary Statistics

Worker Characteristics	**Control**	**Managerial Activity**	**Monitoring**	p-value$^+$
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)	

Notes: Standard deviations 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: Worker Characteristics by Treatment and Control Groups

photo was taken at, and the blurriness of the photo. The managerial activity treatment mentioned the number of the submissions made in a day, and the number that reported the presence of grass. Because of current data limitations we are unable to test the accuracy of grass reporting or the blurriness of photos.[16] In Figures 2, and 3 we compare average changes in task dimensions mentioned in the monitoring treatment and those not mentioned in the treatment group respectively after the beginning of the treatment period by treatment group. The charts presented in Figure 2 demonstrate that the quantity of submissions increased in the monitoring treatment which suggests an improvement in effort. In addition, the charts demonstrate that the increase in how frequently grass was reported before and after treatment began was lower in the monitoring group than in the control or managerial activity groups. This may indicate an increase in the accuracy of grass reporting as a result of the monitoring treatment.[17] However, the Figure also demonstrate that the frequency of night submissions also went up by more in the monitoring group than the control group and by about the same as in the managerial activity group suggesting an economically small decrease in performance on that dimension of output as a result of the both treatments. Night submissions increased significantly in all groups indicating that all workers became less careful over time.

The charts reporting in Figure 3 demonstrate that tree reporting decreased and shrub reporting increased in the monitoring group after treatment which may indicate a trade-off from trees towards shrubs, which can be hard to tell apart, as a result of the monitoring treatment. In contrast, tree reporting increased in the other groups. Shrub reporting also increased in the other groups

[16] The Mechanical Turk data collection effort will allow us to test for changes in these variables.

[17] We will test whether this is the case once data on the accuracy of data classifications has been collected.

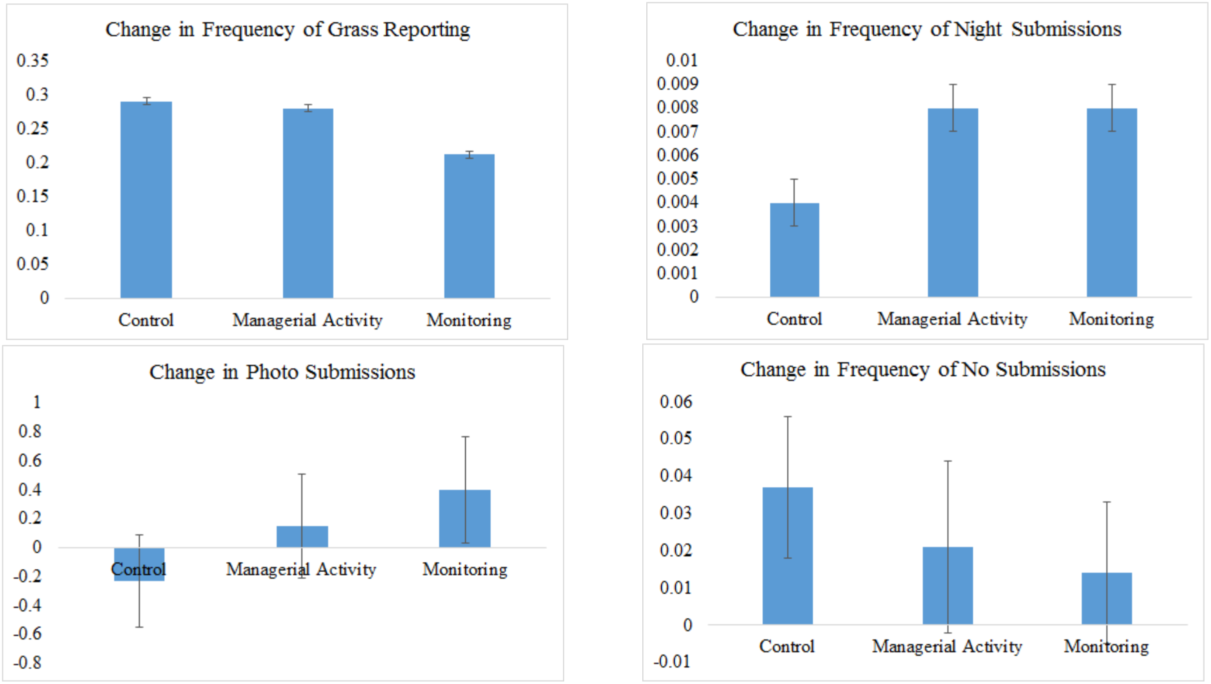

Figure 2: Change in Monitored Characteristics

These figures present the difference between the variable averages before the first treatment was delivered and after the first treatment was delivered across the three groups. Standard errors are presented on each bar.

but by significantly less than in the monitoring group. The Figure also demonstrates that photos seem to decrease in quality over time in all groups, but that those taken by workers in the monitoring group decreased in quality by less (although the difference in quality changes between the managerial activity and monitor groups is not statistically significant). Lastly, the comparison of quit rates across treatment groups demonstrates that worker retention in the monitoring group was significantly higher than in the other groups. Importantly, no workers in the monitoring group left the job before the end of the period whereas 10% in the control and 15% in the managerial activity groups did.[18]

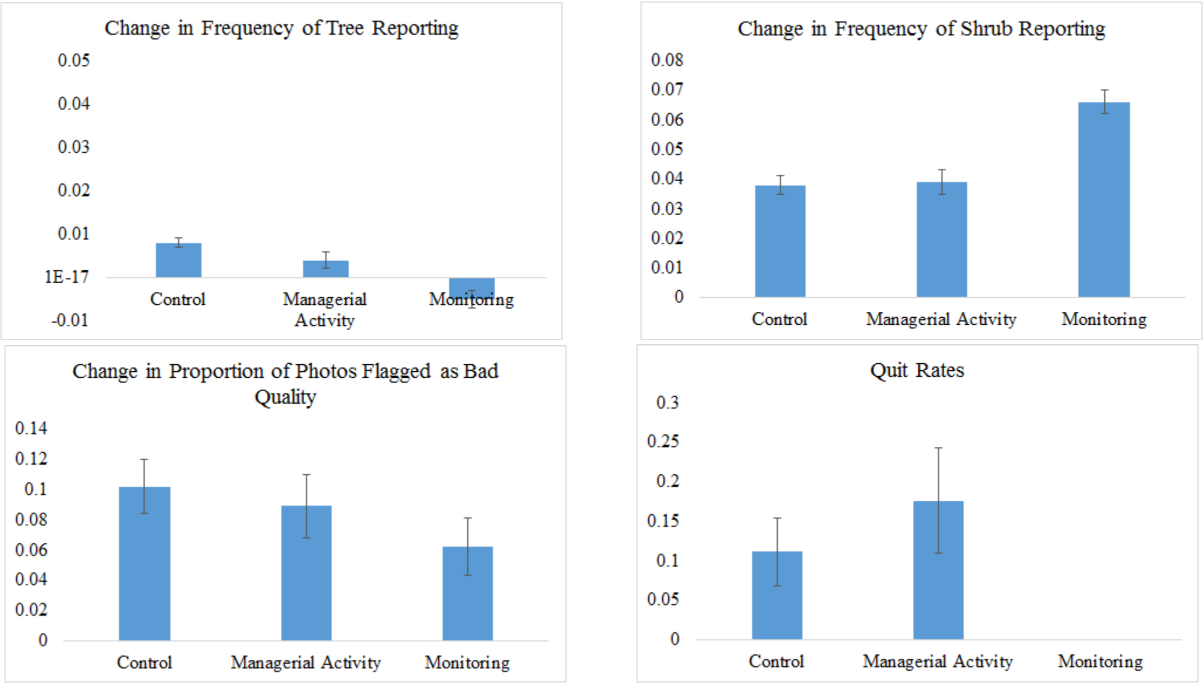

Figure 3: Change in Non-Monitored Characteristics

The figures in the top row present the difference between the variable averages before the first treatment was delivered and after the first treatment was delivered across the three groups. The figure in the bottom row present the proportion of workers who left the job before the end date by group. Standard errors are presented on each bar.

Combined, these mean comparisons suggest that the managerial activity treatment had a small

[18] All worker departures occurred after treatment had begun.

impact on worker performance on some dimensions, and that the monitoring treatment had quite a large impact on worker performance on all dimensions. Moreover, aside from an economically small but statistically significant increase in night submissions, the preliminary evidence presented in Figures 2 and 3 suggest the monitoring treatment improved performance on both monitored and non-monitored dimensions of performance.

4.2 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:

$$\begin{aligned} Y_i = \alpha + \beta_1 ManagerialActivity_i * Treatment_i + \beta_2 Monitoring_i * Treatment_i + \\ \delta Treatment_i + \theta Worker_i + \varepsilon_i, \quad (1) \end{aligned}$$

where Y_i is a characteristics of submission i, $ManagerialActivity_i$ is an indicator for whether submission i came from a worker in the managerial activity group, $Monitoring_i$ is an indicator for whether submission i came from a worker in the monitoring group, $Treatment_i$ is an indicator for whether submission i was submitted after treatment began, and $Worker_i$ is a fixed effect for the worker who submitted i. We also estimate this equation with location fixed effects and treatment group main effects and without worker fixed effects. We estimate a similar equation for number of submissions made per day at the worker-day level of analysis.

To examine whether treatment effects persisted after the treatments stopped, we estimate the following equation:

$$Y_i = \alpha + \beta_1 ManagerialActivity_i * DuringTreatment_i + \beta_2 Monitoring_i * DuringTreatment_i +$$
$$\beta_3 ManagerialActivity_i * PostTreatment_i + \beta_4 Monitoring_i * PostTreatment_i$$
$$+ \delta_1 DuringTreatment_i + \delta_2 PostTreatment_i + \theta Worker_i + \varepsilon_i, \quad (2)$$

where $DuringTreatment_i$ is an indicator for whether or not submission i was made while treatment calls were occurring, where $PostTreatment_i$ is an indicator for whether or not submission i was made after treatment calls had stopped, and the remaining variables are as in equation 1. We report the results of these estimations in the next section.

5 Empirical Results

In this section we report our estimated effects of the managerial activity and monitoring treatments on dimensions of the task that were mentioned as part of the monitoring treatment, and on dimensions that were not. We first report how performance on these dimensions change during the study period following the beginning of treatment. We then explore whether the treatment effects are concentrated within the period during which treatment was occurring, or whether they persist after the treatments have ended.

5.1 Main Results

Table 3 reports the estimated effects of the managerial activity and monitoring treatments on dimensions of the task mentioned in the monitoring treatment calls from the local manager. In particular, we examine how treatment affects the quantity of submissions, the reporting of grass, and whether or not the photos included in submissions were taken at night. The quantity of submissions, and whether or not grass was reported was also mentioned in the calls workers in the managerial activity treatment received from their manager.

In both Panels A and B of Table 3, 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. Estimates change very little across specifications, so we will focus on the specifications that include worker fixed effects going forward.

Panel A reports how the treatment affected the quantity of worker output, and consistent with Figure 2, the results show that the monitoring treatment significantly increased the number of submissions per day, and significantly decreased the number of days with no submissions made. In particular, the coefficient on the interaction between the monitoring treatment indicator and the indicator for whether or not the treatment had begun in column 3 shows that the number of submissions per day increased by about 0.7, 10% of the sample mean, whereas the managerial activity treatment appears to have had no significant impact on submission rates. Similarly, the coefficients presented in column 6 demonstrate that the monitoring treatment decreased the likelihood of not submitting anything on a day by about 8 percentage points, or about 30% relative to the sample mean. Again, the managerial activity treatment did not have an impact on the number of days without submissions for workers.

Panel B reports estimates on how the treatments affected the quality of submissions along the dimensions of the task mentioned in the monitoring treatment. Also consistent with Figure 2, these results show that the monitoring treatment decreased the frequency of grass reporting and increased the frequency of night submissions. The estimates also show that the managerial activity treatment had directionally similar but smaller in magnitude impacts on these dimensions of quality. The interaction coefficients presented in column 3 show that the managerial activity treatment led to a 2.5 percentage point fall in the likelihood of grass being reported, and that the monitoring treatment led to a 9.3 percentage point fall in the frequency of grass being reported. Coefficients presented in column 6 show the managerial activity and monitoring treatments respectively led to a 0.3 percentage point and 0.6 percentage point increase in the frequency of submissions with photos taken at night. However, if we run the analysis with the proportion of submitted photos in a day taken at night, we do not find significant treatment effects, suggesting at least part of this increase is due to the increase in the number of submissions as a result of the treatments.

Combined, these results demonstrate that monitoring improved quantity, the dimension of output that workers were being paid on, but led to a decrease in one dimension of quality, specifically, whether or not a photo was submitted at night. Interestingly, despite the quantity of submissions being mentioned in the managerial activity treatment, workers in that treatment did not change the number of submissions they made. A decrease in the frequency of grass reporting may suggest an increase in submission quality if the decrease reflects greater accuracy. Our current data quality

Panel A: Quantity Outcomes

	Submissions per Day			Day with No Submissions		
	(1)	(2)	(3)	(4)	(5)	(6)
Managerial Activity*Treatment	0.362*	0.357*	0.308	-0.027	-0.027	-0.021
	(0.210)	(0.209)	(0.202)	(0.018)	(0.018)	(0.018)
Monitoring*Treatment	0.684***	0.679***	0.691***	-0.077***	-0.076***	-0.078***
	(0.207)	(0.207)	(0.190)	(0.016)	(0.017)	(0.016)
Treatment	-1.824***	-1.821***	-1.804***	0.203***	0.203***	0.201***
	(0.136)	(0.134)	(0.127)	(0.011)	(0.011)	(0.011)
Managerial Activity	-0.560***	-0.538***		0.076***	0.074***	
	(0.173)	(0.176)		(0.014)	(0.014)	
Monitoring	0.115	0.134		0.017	0.015	
	(0.170)	(0.173)		(0.013)	(0.013)	
Location Fixed Effects	No	Yes	No	No	Yes	No
Worker Fixed Effects	No	No	Yes	No	No	Yes
Observations	16,091	16,091	16,091	16,091	16,091	16,091
R-squared	0.021	0.063	0.395	0.035	0.046	0.353
Mean dep var	6.667	6.667	6.667	0.282	0.282	0.282

Panel B: Quality Outcomes

	Grass Reporting			Night Submission		
	(1)	(2)	(3)	(4)	(5)	(6)
Managerial Activity*Treatment	-0.010	-0.020***	-0.025***	0.004***	0.004***	0.003**
	(0.008)	(0.008)	(0.008)	(0.001)	(0.001)	(0.001)
Monitoring*Treatment	-0.078***	-0.092***	-0.093***	0.004***	0.004***	0.006***
	(0.008)	(0.007)	(0.007)	(0.001)	(0.001)	(0.001)
Treatment	0.291***	0.285***	0.266***	0.004***	0.004***	0.003***
	(0.005)	(0.005)	(0.005)	(0.001)	(0.001)	(0.001)
Managerial Activity	-0.061***	-0.039***		-0.002***	-0.002***	
	(0.006)	(0.006)		(0.001)	(0.001)	
Monitoring	-0.041***	-0.008		-0.003***	-0.003***	
	(0.006)	(0.006)		(0.001)	(0.001)	
Location Fixed Effects	No	Yes	No	No	Yes	No
Worker Fixed Effects	No	No	Yes	No	No	Yes
Observations	107,286	107,286	107,286	107,286	107,286	107,286
R-squared	0.064	0.149	0.399	0.001	0.003	0.035
Mean dep var	0.501	0.501	0.501	0.007	0.007	0.007

Notes: This Table reports estimated effects of treatment on dimensions of the task that were mentioned in the monitoring treatment phone calls. In Panel A, an observation is a worker day. In Panel B, an observation is a submission. Robust standard errors are reported in parentheses. Grass reporting is equal to one if a submission reports that grass is present and zero otherwise. Night submission is equal to one if the photo included in the submission was taken at nigth and zero otherwise. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 3: Effect of Monitoring Activity on Worker Performance: Dimensions Included in Monitoring Activity

collection effort will allow us to test that in the future.

Table 4 reports the estimated effects of the managerial activity and monitoring treatments on dimensions of the task that were *not* mentioned in the monitoring treatment calls from the local manager. In particular, we examine how treatment affects the reporting of trees, the reporting of shrubs, the quality of photos submitted, and whether workers leave the job before the end of the period. As in Table 3, Panel A reports estimates from regressions without fixed effects in columns 1 and 4, with location fixed effects in columns 2 and 5, and with worker fixed effects in columns 3 and 6. In Panel B, estimates from regressions without controls and with location fixed effects are reported.[19]

The estimates presented in Panel A of Table 4 demonstrate that reporting of both trees and shrubs went up in the managerial activity and the monitoring groups once the treatments began. In particular, column 3 shows and 1.5 percentage point increase in tree reporting among workers in the managerial activity group, and a 0.5 percentage increase in tree reporting among workers in the monitoring group. Column 6 shows that shrub reporting went up by about 1.5 percentage points in both the managerial activity and the monitoring groups once treatment began. These magnitudes are quite small relative to the sample means for tree and shrub reporting which were very high even in the absence on the treatments. The results presented in Panel B of Table 4 are consistent with the mean comparisons presented in Figure 3 and show that the monitoring treatment decreased the frequency of poor quality photo submissions, and decreased the number of workers who quit or departed from the job before the end of their contracted term. In contrast, the managerial activity

[19] Worker fixed effects are not included in the analysis of the treatment effects on photo quality because of the small sample size we currently have on this dimension which leaves us with very little within-worker variation. Worker fixed effects are not included in the analysis of the treatment effects on worker departures because all departures occurred after treatment began leaving us with no time variation in departures.

Panel A

	Tree Reporting			Shrub Reporting		
	(1)	(2)	(3)	(4)	(5)	(6)
Managerial Activity*Treatment	0.013***	0.013***	0.015***	0.001	0.003	0.013**
	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)
Monitoring*Treatment	0.004*	0.004	0.005*	0.028***	0.030***	0.015***
	(0.002)	(0.002)	(0.002)	(0.006)	(0.005)	(0.005)
Treatment	-0.009***	-0.008***	-0.004**	0.038***	0.033***	0.038***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)
Managerial Activity	0.004*	0.004		0.032***	0.033***	
	(0.002)	(0.002)		(0.005)	(0.005)	
Monitoring	0.019***	0.018***		-0.009*	-0.006	
	(0.002)	(0.002)		(0.005)	(0.005)	
Location Fixed Effects	No	Yes	No	No	Yes	No
Worker Fixed Effects	No	No	Yes	No	No	Yes
Observations	107,286	107,286	107,286	107,286	107,286	107,286
R-squared	0.003	0.006	0.078	0.007	0.024	0.256
Mean dep var	0.972	0.972	0.972	0.893	0.893	0.893

Panel B

	Poor Quality Photo		Worker Departure	
	(1)	(2)	(3)	(4)
Managerial Activity*Treatment	-0.013	-0.012		
	(0.024)	(0.024)		
Monitoring*Treatment	-0.041*	-0.042*		
	(0.023)	(0.023)		
Treatment	0.102***	0.102***		
	(0.015)	(0.015)		
Managerial Activity	0.021	0.021	0.035***	0.034***
	(0.022)	(0.022)	(0.002)	(0.002)
Monitoring	-0.014	-0.014	-0.040***	-0.041***
	(0.021)	(0.021)	(0.001)	(0.001)
Location Fixed Effects	No	Yes	No	Yes
Observations	23,016	23,016	107,286	107,286
R-squared	0.007	0.007	0.024	0.057
Mean dep var	0.200	0.200	0.0366	0.0366

Notes: This Table reports estimated effects of treatment on dimensions of the task that were not mentioned in the monitoring treatment phone calls. In Panel A and the first two columns of Panel B, an observation is a submission. In columns 3 and 4 of Panel B, an observation is a worker. Robust standard errors are reported in parentheses. Tree reporting is equal to one if a submission reports that trees are present and zero otherwise. Shrubs reporting is equal to one if a submission reports that shrubs are present and zero otherwise. Poor Quality Photo is equal to one if a submission reviewer indicated the photo accompanying a submission was of poor quality 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. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 4: Effect of Monitoring Activity on Worker Performance: Dimensions Not Included in Monitoring Activity

treatment did not have a significant impact on photo quality and increased the number of workers who left the job.

Combined, the results presented in Table 3 show an overall positive affect of the monitoring treatment on worker performance on dimensions not included in what was being monitored. This is consistent with the treatment acting as a signal to workers that their manager was a high productivity manager.

5.2 Persistence of Treatment Effects

To examine whether the treatment effects reported in Tables 2 and 3 are present only while treatment is on-going or whether they also persist after treatment calls have stopped, we divide the period after treatment has begun into a "during treatment" and a "post-treatment" period and examine how treatment effects vary across these periods as described in equation 5. Results from this estimation are reported in Table 5.

Panel A of Table 5 reports the estimates from equation 5 for the dimensions of the task included in the monitoring treatment. The coefficients on the monitoring treatment indicator interactions demonstrate that the treatment effects do persist post-treatment, and in fact appear to get larger across all four outcomes. Interestingly, the managerial activity treatment seems to have had longer term positive impacts on quantity of work completed despite not having any immediate positive effects. Combined, these results suggest that the workers in the treatment groups learn about how to perform their job given what their manager is telling them and that their behavior is reinforced over time.

Panel A: Monitored Dimensions

	Submissions per Day (1)	No Submissions (2)	Grass Reporting (3)	Night Submission (4)
Managerial Activity*	-0.111*	0.019***	-0.014*	0.002
During Treatment	(0.060)	(0.004)	(0.008)	(0.001)
Managerial Activity*	1.297***	-0.025***	0.010	0.004
Post Treatment	(0.102)	(0.007)	(0.011)	(0.003)
Monitoring*	0.726***	-0.091***	-0.060***	0.003***
During Treatment	(0.193)	(0.016)	(0.007)	(0.001)
Monitoring*	1.836***	-0.192***	-0.132***	0.013***
Post Treatment	(0.293)	(0.025)	(0.011)	(0.003)
Observations	16,091	16,091	107,286	107,286
R-squared	0.400	0.363	0.403	0.036
Mean dep var	6.667	0.282	0.628	0.00668

Panel B: Non- Monitored Dimensions

	Tree Reporting (1)	Shrub Reporting (2)	Poor Quality Photo (3)
Managerial Activity*	0.014***	0.007	-0.017
During Treatment	(0.003)	(0.005)	(0.024)
Managerial Activity*	0.007	0.065***	-0.022
Post Treatment	(0.005)	(0.008)	(0.036)
Monitoring*	0.008***	0.016***	-0.032
During Treatment	(0.002)	(0.005)	(0.023)
Monitoring*	-0.011**	0.049***	-0.102***
Post Treatment	(0.005)	(0.008)	(0.035)
Observations	107,286	107,286	23,016
R-squared	0.078	0.256	0.009
Mean dep var	0.972	0.893	0.200

Notes: This Table reports estimated effects of treatment on dimensions of the task that were not mentioned in the monitoring treatment phone calls. In Panel A and the first two columns of Panel B, an observation is a submission. Worker fixed effects are included in the regressions reported in these columns. In column 3 of Panel B, an observation is a worker. The reported coefficients are from a regression that includes location fixed effects. Robust standard errors are reported in parentheses. Tree reporting is equal to one if a submission reports that trees are present and zero otherwise. Shrubs reporting is equal to one if a submission reports that shrubs are present and zero otherwise. Poor Quality Photo is equal to one if a submission reviewer indicated the photo accompanying a submission was of poor quality and zero otherwise. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Effect of Monitoring Activity on Worker Performance During and After Treatment

Panel B of Table 5 reports the estimates from equation 5 for the dimensions of the task not included in the monitoring treatment. We do not include worker departure in this analysis because it does not vary across time. Consistent with Panel A, the coefficient estimates on the monitoring treatment interaction suggest that treatment effects are larger in the longer run and persist even after treatment has ended. There do not appear to be any consistent persistent or immediate treatment effects from the managerial activity treatment.

6 Conclusion

With changes in technology and globalisation, 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. 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 manager productivity through enhanced monitoring on easy to observe dimensions of output.

To test whether increasing monitoring on some dimensions of worker output changes performance on all dimensions, we run a field experiment among workers hired to collect, classify, and transmit data on rangeland conditions in Northern Kenya. Preliminary results from our experiment demonstrate that workers who were randomly assigned to receive additional monitoring from their local manager increased performance on most dimensions of the task discussed during these phones calls. Moreover, their performance on 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 made economically significant changes to their performance. We also find that the treatment effects of enhanced monitoring persist and grow in magnitude after the treatment has ended.

Our results are consistent with workers interpreting increased monitoring on some dimensions of performance being correlated with increased monitoring on other dimensions because active monitors are more productive (e.g. Spence, 1973). Importantly, increased managerial activity without changes in actual monitoring does not achieve this interpretation. Our findings that the effects of the monitoring treatment on performance persist in the longer run and grow in magnitude over time suggests that workers learn how to perform through reinforcement (Börgers and Sarin, 1997; Erev and Roth, 1998). Combined, the results of our experiment demonstrate that relatively low cost increases in how visible monitoring by managers is to workers can lead to economically large changes in performance even without any changes to payment schemes.

In order to better test for quality changes in response to monitoring, we are in the process of collecting additional data on the accuracy of data classifications, and the quality of the photos submitted by workers. This allow us to provide more direct tests of the effects of the monitoring treatment. In addition, we are in the process of investigating how responses to the treatments change depending on how many calls from managers workers had received at a given time, and whether treatment effects stop increasing as the time since the last monitoring call becomes larger. We are also in the process of testing whether treatment effects differ depending on whether workers' received negative or positive feedback on the monitored dimensions of their output.

7 References

Al-Ubaydli, Omar, Steffen Andersen, Uri Gneezy, and John A List, "Carrots that look like sticks: Toward an understanding of multitasking incentive schemes," *Southern Economic Journal*, 2015, *81* (3), 538–561.

Alchian, Armen A and Harold Demsetz, "Production, information costs, and economic organization," *The American Economic Review*, 1972, *62* (5), 777–795.

Baker, George P, "Incentive contracts and performance measurement," *Journal of Political economy*, 1992, pp. 598–614.

_ , Michael C Jensen, and Kevin J Murphy, "Compensation and incentives: Practice vs. theory," *The journal of Finance*, 1988, *43* (3), 593–616.

Bilal, Nejmudin Kedir, Christopher H Herbst, Feng Zhao, Agnes Soucat, and Christophe Lemiere, "Health extension workers in Ethiopia: improved access and coverage for the rural poor," *Yes Africa Can: Success Stiroes from a Dynamic Continent*, 2011, pp. 433–443.

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying, "Does Working from Home Work? Evidence from a Chinese Experiment," *The Quarterly Journal of Economics*, 2015, *165*, 218.

_ , Raffaella Sadun, and John Van Reenen, "Management as a Technology?," *Harvard Business School Strategy Unit Working Paper*, 2016, (16-133).

Börgers, Tilman and Rajiv Sarin, "Learning through reinforcement and replicator dynamics," *Journal of Economic Theory*, 1997, *77* (1), 1–14.

Bresnahan, Timothy F, Erik Brynjolfsson, and Lorin M Hitt, "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *The Quarterly Journal of Economics*, 2002, *117* (1), 339–376.

Courty, Pascal and Gerald Marschke, "An empirical investigation of gaming responses to explicit performance incentives," *Journal of Labor Economics*, 2004, *22* (1), 23–56.

Cragg, Michael, "Performance incentives in the public sector: Evidence from the Job Training Partnership Act," *Journal of Law, Economics, and Organization*, 1997, *13* (1), 147–168.

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

Dumont, Etienne, Bernard Fortin, Nicolas Jacquemet, and Bruce Shearer, "Physicians' multitasking and incentives: Empirical evidence from a natural experiment," *Journal of Health Economics*, 2008, *27* (6), 1436–1450.

Englmaier, Florian, Andreas Roider, and Uwe Sunde, "The Role of Salience in Performance Schemes: Evidence from a Field Experiment," 2012.

Erev, Ido and Alvin E Roth, "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, 1998, pp. 848–881.

Forbes, Silke J, Mara Lederman, and Trevor Tombe, "Quality disclosure programs and internal organizational practices: Evidence from airline flight delays," *American Economic Journal: Microeconomics*, 2015, *7* (2), 1–26.

Frey, Bruno S, "Does monitoring increase work effort? The rivalry with trust and loyalty," *Economic Inquiry*, 1993, *31* (4), 663–670.

Gibbons, Robert, "Incentives in Organizations," *The Journal of Economic Perspectives*, 1998, *12* (4), 115–132.

Holmstrom, Bengt and Paul Milgrom, "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design," *Journal of Law, Economics, & Organization*, 1991, *7*, 24–52.

Hong, Fuhai, Tanjim Hossain, John A List, and Migiwa Tanaka, "Testing the theory of multitasking: Evidence from a natural field experiment in Chinese factories," Technical Report, National Bureau of Economic Research 2013.

Jong, Bart A De and Kurt T Dirks, "Beyond shared perceptions of trust and monitoring in teams: implications of asymmetry and dissensus.," *Journal of Applied Psychology*, 2012, *97* (2), 391.

Kerr, Steven, "On the folly of rewarding A, while hoping for B," *Academy of Management Journal*, 1975, *18* (4), 769–783.

Lazear, Edward P, "Salaries and piece rates," *Journal of business*, 1986, pp. 405–431.

Lu, Susan Feng, "Multitasking, information disclosure, and product quality: Evidence from nursing homes," *Journal of Economics & Management Strategy*, 2012, *21* (3), 673–705.

McPeak, John G and Christopher B Barrett, "Differential risk exposure and stochastic poverty traps among East African pastoralists," *American Journal of Agricultural Economics*, 2001, *83* (3), 674–679.

Mullen, Kathleen J, Richard G Frank, and Meredith B Rosenthal, "Can you get what you pay for? Pay-for-performance and the quality of healthcare providers," *The Rand journal of economics*, 2010, *41* (1), 64–91.

Neuwirth, Benjamin, "Marketing channel strategies in rural emerging markets," *Kellogg School of Management. Available Online*, 2014, *22.*

Prendergast, Canice, "The Tenuous Trade-off between Risk and Incentives," *The Journal of Political Economy*, 2002, *110* (5), 1071–1102.

Reardon, Thomas, C Peter Timmer, Christopher B Barrett, and Julio Berdegué, "The rise of supermarkets in Africa, Asia, and Latin America," *American journal of agricultural economics*, 2003, *85* (5), 1140–1146.

Rubin, Jared, Anya Savikhin Samek, and Roman M Sheremeta, "Incentivizing Quantity and Quality of Output: An Experimental Investigation of the Quantity-Quality Trade-off," *CESR-Schaeffer Working Paper*, 2016, (2016-006).

Slade, Margaret E, "Multitask agency and contract choice: An empirical exploration," *International Economic Review*, 1996, pp. 465–486.

Spence, Michael, "Job market signaling," *The Quarterly Journal of Economics*, 1973, pp. 355–374.

Weisbrod, Burton A, "Rewarding Performance That Is Hard to Measure: The Private Nonprofit Sector," *Science*, 1989, *5*, 244.