informs
http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 12, December 2017, pp. 4061–4080
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Role of Communication of Performance Schemes: Evidence from a Field Experiment

Florian Englmaier,[a] **Andreas Roider,**[b] **Uwe Sunde**[a]

[a] Department of Economics, LMU Munich, 80539 Munich, Germany; [b] Department of Economics, University of Regensburg, 93040 Regensburg, Germany

Contact: englmaier@lmu.de (FE); andreas.roider@ur.de (AR); uwe.sunde@lmu.de (US)

Received: April 11, 2014
Revised: August 7, 2015; December 18, 2015; March 7, 2016
Accepted: April 27, 2016
Published Online in Articles in Advance: October 13, 2016

https://doi.org/10.1287/mnsc.2016.2559

Abstract. In corporate practice, incentive schemes are often complicated even for simple tasks. Hence, the way they are communicated might matter. In a natural field experiment, we study a minimally invasive change in the communication of a well-established incentive scheme—a reminder regarding the piece rate at the beginning of the shift. The experiment was conducted in a large firm where experienced managers work in a team production setting and where incentives for both quantity and quality of output are provided. While the treatment conveyed no additional material information and left the incentive system unchanged, it had significant positive effects on quantity and on managers' compensation. These effects are economically sizable and robust to alternative empirical specifications. We consider various potential mechanisms, but our preferred explanation is that the treatment raised the salience of incentives.

History: Accepted by John List, behavioral economics.

Funding: The authors gratefully acknowledge financial support from the German Science Foundation [DFG, SFB/TR-15] and the European Commission [SCIFI-GLOW, Contract SSH7-CT-2008-217436].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2016.2559.

Keywords: incentives • attention • salience • communication • natural field experiment

1. Introduction

1.1. Motivation

Incentive schemes designed to enhance workforce performance are a key element of corporate personnel policy and have received considerable attention in academic research. In particular, a growing empirical literature studies the effectiveness of incentive schemes and sheds light on important behavioral aspects. However, most of this evidence is based on fairly straightforward incentive schemes and drastic interventions, as, for example, in Lazear's (2000) seminal Safelite study.[1] In corporate practice, however, many incentive schemes (and the organizational structures in which they are implemented) are complex even for simple tasks. While the focus in the theoretical literature has been on the optimal design of incentive schemes, the empirical question whether existing incentives in fact lead to optimal performance by the workforce has not received much attention.

In this paper we investigate whether a minimally invasive change in the way an elaborate and well-established incentive scheme is communicated has measurable effects. The evidence stems from a natural field experiment conducted in a large firm. Over the years the firm had developed an incentive scheme for managers, who were each responsible for a team in a complex multidimensional production process where both quantity and quality of output are important. While keeping the material incentive structure—under which managers had worked for several years—unchanged, the experiment varied how critical information about the existing performance scheme was communicated. In particular, the only change that happened in the randomized intervention was that a reminder about the prevailing piece rate was posted at the beginning of the shift; a crucial piece of information that managers already had absent the treatment.

Given the experience of managers (who, on average, have worked for the firm for several years) and given that incentive pay constitutes a substantial fraction of their income, under the null hypothesis, the performance and the earnings of these managers should not be affected by a reminder about the piece rate. That is, this variation in the way incentives are communicated should have no impact on performance. The results document, however, that the intervention had economically and statistically significant effects on outcomes in a real-world production environment with substantial monetary incentives, even though the existing incentive system remained unchanged throughout the treatment and control periods and across treatment and control teams.

1.2. The Firm

The experiment was conducted in cooperation with a large European agricultural firm whose main product is lettuce. The harvesting of lettuce is done by teams of harvest workers that work together on a harvest machine and perform various tasks. Each of the teams (which work on different fields in the same region) is led by a manager who is responsible for this one team only and serves as the crucial link between the firm and workers. The respective manager oversees the operations of the team. In particular, the manager is responsible for the harvest performance and takes the relevant operative decisions (e.g., the speed of the harvest machine, the matching of workers to tasks, and the training of incoming workers) for his entire team. The manager of the team also communicates the output requirements as well as the incentive structure to the workers (see Section 3.1 for details).

The firm cares both about the harvested quantity and quality because it faces severe contractual penalties for inferior quality delivered to large supermarket chains. Accordingly, incentives are set twofold. Quantity incentives are provided via a piece rate. This rate is determined ex-ante by the firm's headquarters for each team and shift separately to set incentives, but at the same time to adjust for varying conditions with respect to weather, field, crop, and demand. Quality incentives are provided via deductions from team pay for deficient quality as well as through a daily tournament scheme across teams in which the teams delivering the highest qualities win (potentially substantial) monetary prizes. Quality is measured by regular predelivery quality checks. The incentive structure is explained in more detail in Section 3.2.

1.3. The Experiment

The controlled experimental intervention varied the communication of incentives, while the actual monetary incentive system remained unchanged.[2] In the preintervention (control) situation, managers were informed about the piece rate pertaining to their team on the respective day when each of them reported to the firm's headquarters before the start of his shift. However, there was neither monitoring whether managers acknowledged this crucial variable, nor whether they communicated it to their workers.[3] In the experimental treatment, the firm changed the communication of ex-ante determined piece rates for one month for a randomly selected group of managers and their respective teams. The intervention ensured that treated managers and workers received this information as both managers and their teams were explicitly briefed, and a note stating the current piece rate was posted on the harvest machine, visible to the entire team, at the beginning of the respective shift. The experimental intervention is described in more detail in Section 3.3, and Section 4 contains a description of the data and the estimation strategy.

1.4. Results

We find that the intervention had economically sizable effects. It significantly increased output (by about 3.4%–3.8%) and had a negative (but not robustly significant) effect on quality (by about 2.1%–5.4%).[4] Moreover, the intervention significantly increased manager daily compensation (by about 4.1%–4.8%). These findings survive a host of robustness checks like alternative assumptions on clustering (including bootstrapped standard errors that allow for clustering on the day and team level), varying sets of control variables (like excluding the gross piece rate), and various specifications of the sample period; see Sections 5.1 and 5.2.

Investigating potential channels for the treatment effects sheds light on how managers change their behavior in response to the intervention (see Section 5.3). In particular, managers in treated teams start to assign a larger fraction of their workers to the task of cutting the lettuce (the central and most demanding task in harvesting lettuce), which approximately accounts for the treatment effect on quantity. Moreover, the change in the communication of quantity incentives appears to lead to behavior that is more finely adjusted to material incentives (which vary across days and teams). In contrast, worker behavior and compensation are not affected by the intervention. We also find evidence that the treatment effects take time to build up over the course of the treatment month.

1.5. Interpretation

In Section 6, we take a closer look at the results to explore why our intervention might have led to the substantial treatment effects described in Section 1.4. While alternative mechanisms (such as that either managers or workers infer additional information from the treatment or a Hawthorne effect) might in general play a role, we argue that they are either inconsistent with some of our findings or implausible in the current setting. Our preferred interpretation, which seems consistent with all of our empirical findings, is that the intervention increased the salience of the incentive structure to managers: While the piece rate directly influences their pay, managers face a variety of tasks, and even in the present (relatively straightforward) production setting the incentive system is complex (with various incentive instruments for quantity and quality). In addition, beyond the immediate supervision and direction of workers, managers also have to decide on the allocation of workers to tasks and to train incoming workers. Our findings are consistent with the interpretation that the experimental intervention refocused managers' attention on the incentive system, thereby allowing them to obtain a higher payoff.

The remainder of this paper is structured as follows. In Section 2 we discuss the related literature. Section 3 introduces the firm where we conducted the field experiment, its team production technology and incentive system, and describes the experimental treatment. Section 4 describes our data, formulates our predictions, and explains our empirical specification. Section 5 presents our results and various robustness checks. Section 6 discusses potential mechanisms. Section 7 summarizes the results and discusses their contribution given an arguably low prior for finding an effect of a slight change in the communication of incentives.

2. Related Literature

In light of the potential interpretation of our findings as a consequence of changes in the relative salience of incentives, this study contributes to a recent empirical literature documenting effects of inattention. So far, this literature has mainly focused on consumption choices and personal finance; for a survey, see DellaVigna (2009). For example, various authors consider online auctions and show that bidders are inattentive to relevant information. In particular, in field experiments, Hossain and Morgan (2006) and Brown et al. (2010) document that, if the salience of shipping costs is low (e.g., because they are stated separately from the price), shipping costs are not fully incorporated into buyers' bidding decisions. Lee and Malmendier (2011) show that bidders frequently fail to exercise available (advantageous) "buy-it-now" options. In a similar vein, the degree of salience of taxes appears to affect consumption behavior. For example, Chetty et al. (2009) conduct a field experiment at a grocery store and find that posting tax-inclusive prices reduces demand. Finkelstein (2009) shows that reduced salience of road tolls (caused by the introduction of electronic toll collection systems) leads to higher tolls. There is also evidence that consumers do not fully appreciate the continuity of price or quality measures and instead frequently focus on a coarser grid of (focal) values when making decisions; see, for example, Lacetera et al. (2012) and Pope (2009). In the realm of personal finance, various studies have documented that behavior varies systematically with the way institutional features are communicated. For example, Karlan et al. (2016) conduct a field experiment documenting that reminders that are sent to savings account holders are more effective in changing savings behavior when they increase the salience of specific expenditures. In another field study, Stango and Zinman (2014) manipulate the salience of checking overdraft fees by injecting overdraft-related questions into surveys and find that increased salience has the immediate effect of reducing the likelihood of incurring a fee in the current month. Moreover, taking part in multiple overdraft-related surveys seems to build a "stock" of attention that reduces overdrafts for up to two years. In contrast to the studies on consumption choices and personal finance, our paper considers incentive provision within a firm and documents in a natural field experiment that varying the communication of certain aspects of the incentive system substantially affects performance even in a context with experienced managers.

Another related strand of recent papers investigates the consequences of variation in the information about incentives that is provided to the workforce (while holding the monetary incentive system fixed).[5] For example, Blanes-i-Vidal and Nossol (2011) consider a setting where workers are paid piece rates and where management begins to reveal the relative position of workers in the pay and productivity distribution. It turns out that this additional information about relative performance leads to substantial and lasting increases in productivity (e.g., due to social comparison processes), even though the material incentives have not changed. In a field experiment, Barankay (2012) finds strong negative effects of rank information on performance among male employees. Bandiera et al. (2013) show that introducing performance feedback, without changing incentives themselves, has measurable effects on output. In their study, feedback information generates incentives to change the (endogenous) team composition by making clear the benefits of assortative matching into teams by ability. Finally, Hossain and List (2012) report on a field experiment studying the effects of the introduction of conditional incentives framed as either "losses" or "gains" in a Chinese high-tech manufacturing facility. While both are shown to increase productivity, performance persistently responds stronger to incentives that are framed as losses than to identical incentives that are framed as gains. Our paper complements these studies because in our natural field experiment managers had access to the same information both in the control period and in the treatment period, and our intervention only changed the way this information was communicated, while keeping the framing of incentives fixed. This allows us to focus on the pure effect of the intervention regarding the communication of the existing incentive system.[6]

Our results speak also to the empirical analysis of the multitasking problem. Going back to Holmstrom and Milgrom (1991), the theory suggests that effort substitution between two tasks arises if increasing effort on one task increases marginal costs of effort in the other task; this need not be the case if marginal costs are mutually unaffected. In many real-world settings, like our quantity-quality trade-off, a complementarity in marginal costs appears plausible. Despite the rather clear-cut theoretical predictions, the empirical support for effort substitution has been somewhat mixed. On the one hand, Paarsch and Shearer (2000), for the case of tree planting, Johnson et al. (2015), for the case of Chilean bus drivers, Dumont et al. (2008), for the case of physicians, and Hong et al. (2013), for the case of Chinese factory workers, find evidence for increased effort in the quantity and decreased effort in the quality dimension after a piece rate incentive scheme was introduced. On the other hand, the field experiments of Lazear (2000), Shearer (2004), Bandiera et al. (2005), and Al-Ubaydli et al. (2015) do not find that the quality of work is affected by incentives for the quantity of production. While in our setting there was no material change in the incentive system, based on our preferred interpretation, the intervention might have led the manager to pay closer attention to the piece rate and thereby might have led to effectively stronger quantity incentives. The results indicate a significantly positive effect on quantity, and systematically negative point estimates for the effect on quality, which, however, are not always significant.

3. Setting and Experimental Design

3.1. Technology and Workforce

The firm where the field experiment was conducted is a large European agricultural producer that mainly grows vegetables. For the current study, we use data on all teams that harvest (a certain variety of) lettuce, the firm's main product. The harvest season starts in May and ends in November, with June–September being the peak harvest season.

3.1.1. Harvest Teams Consisting of a Manager and (Temporary) Workers. The harvesting is done using a team technology, where on every day of the week around 10 teams independently harvest lettuce in shifts on different fields in the same geographical region. Teams are, in general, too far apart from each other to directly communicate during a given shift. Each team uses a separate harvest machine (that economizes the entire harvest process) and typically consists of a dedicated manager (who is a long-term employee of the firm) and more than 30 (temporary) workers, who fulfill various tasks within the harvest team. In particular, on average 10–12 cutters (standing behind the harvest machine) do the actual harvesting: they cut the lettuce, put it in a plastic bag, and place it on a conveyor belt, which is attached to the machine. From there, packers (who sit behind the belt) pack the lettuce in crates. Crate-staplers subsequently transport the crates to the center of the harvest machine and put them on palettes (which are then wrapped with foil and put onto a trailer in front of the harvest machine by the stretchers). The trailer and the harvest machine are pulled by a tractor.

The manager as the leader of his respective harvest team identifies a team in our data. As the other members of the respective harvest team are temporary workers, team composition varies over the course of the harvest season (as will be discussed in more detail in Section 3.1.2). Each manager has a variety of responsibilities. He is the link between the firm and workers, communicates details about the incentive structure to workers, is responsible for training of incoming workers, and takes all relevant operative decisions on the field. For example, within his team he decides on the allocation of workers to various tasks (i.e., assigns them to be cutters, packers, crate-staplers, or stretchers). Also, the manager sets the speed of the harvest machine (and thereby implicitly decides how much lettuce is worth harvesting on a given field on a given day). Ultimately, the manager is responsible for the entire performance of his team in terms of quantity and quality of lettuce harvested.

3.1.2. Workforce, Allocation of Workers to Teams, and Training. Unlike managers, workers generally only stay with the firm for spells of six to eight weeks of a harvest season. However, it is not uncommon for workers to return over multiple years in a row for these short spells. Importantly, before starting to harvest, all incoming workers receive training that introduces them to the production technology, the various tasks, and the pertaining incentive structure.[7] Workers are mostly from Eastern Europe (mainly from Poland, Romania, or Ukraine).[8] In general, these temporary workers are recruited in their home towns, for example, upon recommendation by workers from previous years. Arrivals and departures at the firm site are organized by the firm in batches of bus loads to make travel cost-efficient. During their spells, the workers live on the farm at centrally provided lodging sites. Incoming workers are allocated to managers by the firm's headquarters (and not by the managers themselves) and stay with their respective manager's team for their entire spell at the firm. According to the firm's headquarters, which team a given worker will join is not a conscious decision but basically random and driven by current departures of workers from the firm site (i.e., in which teams there are openings). Importantly, this implies that from the perspective of the current paper the allocation of workers to teams can be seen as exogenous.

3.2. Incentive System: Concerns for Quantity and Quality

3.2.1. Overview. The firm cares about both the quantity—higher output increases revenue—and quality—severe contractual penalties would result from inferior quality delivered to large supermarket chains—of the lettuce harvested. As a consequence, the firm maintains an elaborate incentive system to provide quantity and quality incentives to both managers and workers. This incentive system has been in place for several years prior to our experiment.

In the following, we describe the remuneration of workers. While managers face a very similar pay structure, we postpone the details pertaining to their remuneration to the end of this subsection. Importantly, note that for managers firing or promotions, i.e., career concerns, do not appear to be an issue as so far neither of the two has happened at the firm. In a similar vein, according to the firm, firing of temporary workers due to a lack in performance is an extremely rare event (which does not seem to be surprising given the simplicity of the tasks and the attractiveness of the job in terms of potential earnings).[9]

3.2.2. Quantity Incentives. Incentives for quantity are provided through piece rates for the amount of lettuce harvested. Piece rates (which are team-day specific) are set by the firm's headquarters each day before a given shift begins. Importantly, in the present firm, piece rates fulfill a twofold purpose. In addition to providing incentives, the firm has to ensure that workers obtain an average hourly wage above the legal minimum. Consequently, adjustments are made to the piece rate to account for varying harvesting conditions, such as the condition of the field (e.g., soil or field size), crop (e.g., size of the lettuce heads, maturity, or potential damages), as well as weather conditions.[10] For a given team and day, the average hourly pay of a worker from quantity incentives is given by the piece rate (in terms of lettuce heads) times the total number of pieces harvested divided by the total number of work hours.[11]

3.2.3. Quality Incentives. The quality of a given team's output is measured by a one-dimensional index (the so-called quality (malus) points), where a higher number of points reflects worse quality of the product (e.g., damaged leaves or brown stains) or of the harvest process (e.g., compliance with work hygiene). The quality assessment is conducted by designated quality control staff at the team-day level, postharvest and predelivery at the firm's warehouse, as well as on site, where mobile quality control staff visits each team during each shift, as will be discussed in more detail in Section 3.3.

The number of quality (malus) points affects workers' pay in two ways. First, deficient quality reduces a worker's pay directly through a deduction to the payment from quantity incentives, where the deduction (per piece harvested) is proportional to the assigned number of quality points. Second, there is a daily quality tournament. In particular, the deductions are paid into a pool. At the end of the harvest day, this pool is distributed through a tournament among all teams active on the respective day, where the teams with the best quality performance (the lowest number of quality (malus) points) receive percentage-shares of the pool (i.e., prizes) that are decreasing in their rank in the tournament. The fixed-percentage distribution scheme is known to all managers and workers. However, the absolute size of the prizes is determined endogenously by the size of the pool, i.e., by the quality performance of all teams on the respective harvest day. Any payout from the quality tournament is distributed equally among the workers of a given team.

3.2.4. Specifics of Manager Remuneration. Analogous to workers (see Sections 3.2.2 and 3.2.3), managers receive quantity incentives (through the piece rate) and quality incentives (through deductions proportional to the number of quality points and participation in daily quality tournaments). Compared to workers, quality incentives do, however, receive a larger weight in managers' remuneration. In addition, they receive a base wage, which results from collective bargaining agreements, and they participate in the firm's "profit center harvest" (where a certain percentage of the respective harvest day's profits is distributed among all managers active on a given day). This latter component is meant to provide managers with incentives for good maintenance of machinery, economical usage of material, etc.

3.3. Experimental Implementation

The present paper is the result of an intense interaction with the firm, whose board could be convinced of the merit of conducting a controlled experiment to evaluate the effectiveness of the incentive system in place. As the incentive system had been developed over years, the firm was convinced of its effectiveness. Based on the discussions in Sections 1 and 2, we suspected, however, that small interventions in how incentives are communicated might matter. As our proposed experimental design was not considered to pose a danger of disrupting the production process, it was approved. Note that neither managers nor workers were informed that they were exposed to an experimental intervention.

3.3.1. Randomization. The experimental treatment was conducted between August 1 and August 31, 2008. The treated population consisted of five managers and their respective teams, who were randomly drawn from the set of managers. The control population consisted of the remaining five managers and their respective teams. As explained in detail in Section 3.3.2, the experimental intervention changed the way quantity incentives were communicated to treated teams, while keeping the actual monetary incentive system fixed.

3.3.2. Implementation. Outside of August 2008, when managers reported to the firm's headquarters before the beginning of their shift, they were informed about their respective piece rate (as well as about other relevant aspects, such as the relevant field and the length of their shift). However, there was no monitoring as to whether managers actually acknowledged this information or whether they communicated the piece rate to their workers. During the treatment in August 2008, managers still received the relevant information before the beginning of their shift at the firm's headquarters. In addition, our intervention made sure that managers and workers acknowledged the information about the pertaining piece rate. In particular, the intervention was conducted by the firm's mobile quality control staff, whose regular task is to monitor the production process on-site at the various harvest machines. To this end, the quality control team regularly visits each team at the beginning of its respective shift. Hence, the mobile quality control staff's visit is not per se perceived as an unexpected intervention. During the treatment in August, before the shift began, the mobile quality control staff briefed the respective (treated) manager explicitly about the pertaining piece rate by posting a note stating the piece rate on the harvest machine where it was visible to the entire team. The mobile quality control staff was instructed to put up the note and not to comment proactively on the issue. In case someone asked they were scripted to say, "Headquarters told me to put this up." In case of further enquiries, they were instructed to respond, "This is information that you already know," and not to respond to further follow-up questions. Beyond this (and its usual tasks), the mobile quality control staff did not intervene in the production process. The mobile quality control staff was ordered by the firm to implement the intervention without being told that it was actually an experiment. In communication with the firm, it was reported that there were no particular inquiries regarding the intervention.[12]

The coefficient of interest is the effect of the treatment on the treated teams compared to the respective outcomes in control teams, i.e., the treatment effect on the treated.[13]

4. Data, Predictions, and Empirical Specification

4.1. Data

Our analysis relies on the personnel and performance records for all managers and workers for the harvest season 2008. The harvest season 2008 lasted from May 25 to November 6, 2008, and our unit of observation are data on the team-day level. As discussed in Section 3.1, a team is identified by its respective manager (because a given team's stock of temporary workers fluctuates, where leaving workers are replaced by current arrivals). In the analysis, we only consider teams that work on performance pay, eliminating approximately 15% of team-day observations in which teams are working on fixed wages due to bad conditions or other reasons unrelated to the intervention.[14] This yields 1,182 team-day observations for 5 treated teams (534 observations) and 5 control teams (648 observations), whereas in the treatment period of August 2008, for both treated teams and control teams we have 107 observations each. All variables (except binary indicator variables and fractions) have been standardized to a mean of zero and a standard deviation of one to protect confidential firm information.[15]

As a preliminary step, we investigate whether the randomization of treated and control teams was successful, and we look for preexisting trends in the data. As displayed by Table 1, there is basically no evidence for systematic differences in observable characteristics between treated teams and control teams prior to the treatment (where we focus on July 2008, i.e., the month before the treatment). As all variables reported in Table 1 have been standardized to a mean of zero and a standard deviation of one (on the entire sample period), the differences in means in Table 1 can be interpreted in terms of standard deviations (over the entire harvest season). We report t-test statistics to check

Table 1. Balancing Table

	Mean			p-value	
	Control teams	Treated teams	Difference	t-test	KS-test
Piece rate	0.610	0.782	−0.172	0.18	0.83
Total shift length	0.406	0.492	−0.086	0.54	0.44
Break time	0.233	0.491	−0.258	0.10	0.66
Team size	−0.009	−0.175	0.166	0.27	0.32
Workers' average age	−0.105	−0.039	−0.066	0.66	0.01
Workers' average total tenure	0.343	0.244	0.099	0.43	0.41
Workers' average current tenure	0.707	0.798	0.091	0.05	0.06

Notes. Observations are on the team-day level, where, for confidentiality reasons, all variables (except for dummy variables and fractions) are standardized to a mean of zero and a standard deviation of one on the entire sample of 1,182 observations as in the main analysis. *Workers' average total tenure* relates to the number of seasons workers have worked for the firm (including the current one). *Workers' average current tenure* relates to the number of days workers have worked for the firm in the current season. The table considers observations from July 2008 only (i.e., the month prior to the treatment). This yields 99 (117) observations for treated teams (control teams), where the p-values refer to t-tests for the null of equality of means and Kolmogorov–Smirnov (KS) tests for the null of equality of distributions in the two groups, respectively.

for differences in means and Kolmogorow–Smirnov test statistics for differences in the distributions. While workers in treated teams exhibit a slightly shorter tenure in the current season (0.091 standard deviations below that of control teams), this might be driven by differences in the timing of arrival and departure of temporary workers, which are randomly allocated to the various teams. In the nonstandardized data, this difference in means corresponds to less than three days. Similarly, the difference in break time of 0.26 standard deviations (which is not significant at the 5% level and where the Kolmogorow–Smirnov test does not show distributional differences) translates into a difference of less than five minutes.[16] For workers' average age, the t-test indicates no difference in means while the Kolmogorow–Smirnov test indicates distributional differences. At the mean, the difference corresponds to four months of age difference. However, this difference is unlikely to affect performance given the simplicity of the tasks and the required (short) training period. Overall, it appears as if treated teams and control teams did not differ systematically with respect to personal characteristics and experience of workers or with respect to inputs before treatment. Complementing this, we find no difference between control and treated teams in performance before the treatment when controlling for team characteristics (see Table 7 and the corresponding discussion in Section 6). Moreover, there is no evidence that any potential differences between treated teams and control teams along these dimensions varied systematically across control and treatment periods. Importantly, incentives in terms of the piece rate set by the firm do not differ systematically between teams receiving the treatment and teams that are in the control group, neither before nor during the treatment. The respective p-values from t-tests are 0.13 over the entire season, 0.18 for the month before the treatment period (July), and 0.78 (0.90) for August (after August 1), respectively.

The raw data reveal two observations that are relevant for the empirical analysis (see Figure R.1 in the online appendix). First, there do not seem to be differential preexisting trends between treated teams and control teams in any of the outcome variables. Second, performance along all three of these dimensions is subject to pronounced day-to-day fluctuations (e.g., due to substantial day-to-day changes in harvesting conditions). This suggests that a multivariate regression approach that accounts for systematic variation in observable team and day characteristics is more appropriate for the identification of the treatment effect than a comparison of unconditional means. In the empirical analysis in Sections 5 and 6, a full set of control variables will be used to identify the effects of interest.

4.2. Predictions

Given that the incentive system as described in Section 3.2 has been in place for several years, that managers are experienced, and that their pay (as well as workers' pay) is highly incentive-based, standard agency theory suggests that managers maximize against the incentive scheme. Hence, our null hypothesis is that the slight experimental variation of the communication of the piece rate affects neither managers' behavior nor their pay.

However, when devising the experiment, we deemed it possible that a change in the communication of the piece rate might affect managers' behavior for (at least) the following two reasons. First, as discussed in Section 2, there is evidence from a variety of domains (but, so far, not incentive provision in firms) that variations in salience might affect behavior. Second, while the piece rate directly influences managers' pay, the team managers face a variety of tasks, and even in the present (relatively straightforward) production setting the incentive system managers face is complex. In particular, there are not only quantity incentives through the piece rate, but managers also need to be concerned about quality (incentivized through deductions from pay and daily quality tournaments). In addition, beyond the immediate supervision and direction of workers, managers have to decide on the allocation of workers to tasks and to train incoming workers. Consequently, it might be that our intervention with respect to the communication of (quantity) incentives brings about a change in treated managers' behavior.

When testing the null hypothesis, a subtlety arises with respect to the definition of the control period. The part of the harvest season before the treatment (i.e., from May 25 to July 31, 2008) clearly serves as control. However, given that the treatment constitutes a change in the communication of incentives, and thus potentially in the manager's perception of the incentive system, it might be debatable how to deal with the part of the harvest season after the end of the treatment (i.e., from September 1 to November 6, 2008). Strictly speaking, the experimental treatment was only applied in August 2008, implying that the period from September 1 onward should be viewed as a control period. However, it might be that exposure to the treatment led managers to change their behavior even after August 2008 (when they were no longer explicitly reminded of daily incentives). In this case, the entire remainder of the harvest season after August 1 could be viewed as the treatment period. To deal with this issue, in our regression analysis, we provide results for both interpretations. In Section 5.2 we provide a more detailed discussion and show that our results also hold when estimating an empirical model that accounts for the posttreatment period separately, or when dropping the entire period after August 31.

4.3. Empirical Specification

The analysis builds on team-day observations, where teams are identified through their respective manager. In establishing our results, we begin by illustrating the treatment effects graphically in the data. To this end, we plot raw means of the relevant outcome variables (quantity, quality, and manager's pay) across treatment status and period. However, there are substantial day-to-day fluctuations in the variables of interest, for example, due to changes in environmental conditions (see Figure R.1 in the online appendix). Hence, the variation of the outcome variables is large, which suggests that a regression-based identification of the treatment effect is more appropriate. The empirical analysis is based on an estimation model that identifies the treatment effects of the experimental intervention by ways of a difference-in-differences approach, where outcomes for treated teams before and under the treatment are compared to those of control teams that did not receive the treatment. The empirical specification conditions on a rich set of controls to avoid spurious results driven by systematic heterogeneity. In particular, our empirical model is given by

$$\begin{aligned} Y_{it} &= \alpha + \beta \cdot \mathcal{I}_{\text{Treated team}} \cdot \mathcal{I}_{\text{Treatment period}} \\ &\quad + \gamma_i \cdot T_i + \delta_t \cdot D_t + \rho \cdot X_{it} + \varepsilon_{it}, \end{aligned} \tag{1}$$

where i and t denote team and day, respectively; Y_{it} is the respective outcome variable; $\mathcal{I}_{\text{Treated team}}$ ($\mathcal{I}_{\text{Treatment period}}$) is a binary indicator that is equal to one for treated teams (during the treatment period) and zero otherwise; T_i is a binary team indicator (where a team is identified by its manager); D_t is a binary day indicator; X_{it} is a vector of controls on the team-day level; and ε_{it} is an error term. The coefficients to be estimated are α, β, γ_i, δ_t, and ρ, where β is the coefficient of interest as it reflects the effect of the treatment on the treated.

Note that the controls X_{it} and the (manager) fixed effects T_i capture distinct aspects. On the one hand, T_i captures persistent differences that are based in the person of the manager (e.g., management style or authority). On the other hand, the (team) controls reflect properties of the team that, for a given manager, might vary from day to day, for example, team composition due to departure and arrival of new workers. Typically, on any given day, these controls will also vary across teams. In particular, the covariates X_{it} control for factors that might affect performance but that in our setting can be viewed as exogenous from the respective manager's perspective. This includes the material incentives in terms of the piece rate set by the firm before the respective shift. To account for potential productivity differences across teams, controls also include information regarding (i) team composition in the form of the average age of workers and the average tenure of workers (in terms of both the number of days worked in the current season and the overall number of seasons worked for the firm),[17] (ii) the labor force at the disposal of the manager on a given day (in terms of team size, i.e., the number of workers, and of total work hours),[18] and (iii) a binary indicator that is equal to 1 if the team went on to work on a second shift on the respective day (see Endnote 14).

We estimate different versions of model (1). As main specification we present ordinary least squares (OLS) estimates with robust standard errors, where we allow for correlation of errors within a harvest day, for example, due to weather effects or the daily quality tournaments. In the light of the raw data, this clustering appears to be an appropriate assumption as it accounts for the most important source of unobserved heterogeneity, while maintaining a sufficient number of independent observations (see, e.g., Cameron et al. 2011), and it follows comparable approaches in the literature (see, e.g., Blanes-i-Vidal and Nossol 2011). As alternative, we also report 95% confidence intervals based on standard errors that allow for clustering on the day and team level, where the team dimension is bootstrapped (Cameron et al. 2008). In the robustness section, we discuss results obtained under alternative assumptions about clustering.

5. Results

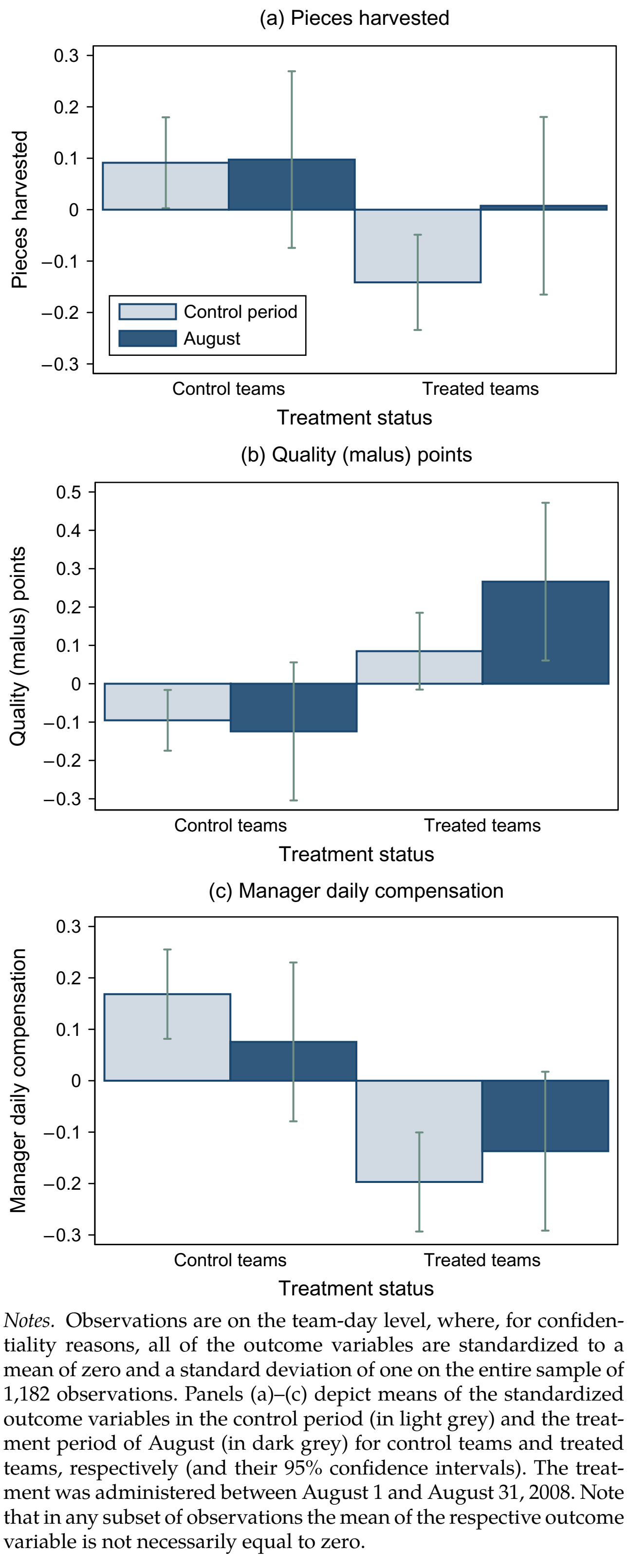
In Section 5.1 we establish the effects of the intervention on the main outcome variables (i.e., quantity and quality of output and manager daily compensation), and in Section 5.2 we show that the results are robust to alternative specifications. Then, in Section 5.3, we document various behavioral responses that shed light on how the treatment effects were moderated.

5.1. Main Results: Treatment Effects on Outcome Variables

5.1.1. Quantity. In Figure 1(a) we compare the raw means of quantity harvested in the control period and the treatment period (of August 2008) for control teams and treated teams, respectively. Recall that all outcome variables have been standardized to a mean of zero and a standard deviation of one to protect confidential firm information, which explains why in Figure 1 the outcome variables may take on negative values.[19]

While treated teams display a somewhat lower quantity prior to treatment when compared to control teams, their quantity output goes up in August, while that of the control teams remains virtually unaffected by the treatment. In particular, in the control period the raw means are 0.091 and −0.141 for control teams and treated teams, respectively, and hence there is a difference of 0.232. In the treatment period of August, the raw means are 0.097 and 0.008; implying a difference of 0.089 between control teams and treated teams.

Figure 1. (Color online) Comparison of Means by Treatment Status and Treatment Period



Notes. Observations are on the team-day level, where, for confidentiality reasons, all of the outcome variables are standardized to a mean of zero and a standard deviation of one on the entire sample of 1,182 observations. Panels (a)–(c) depict means of the standardized outcome variables in the control period (in light grey) and the treatment period of August (in dark grey) for control teams and treated teams, respectively (and their 95% confidence intervals). The treatment was administered between August 1 and August 31, 2008. Note that in any subset of observations the mean of the respective outcome variable is not necessarily equal to zero.

The treatment effect, in terms of the difference of differences, thus amounts to 0.143 standard deviations. This corresponds to a relative daily increase in pieces harvested by treated teams by about 4%. As is evident from Figure 1(a), variation in the raw data is, however, large.[20]

Consequently, on the basis of the difference-in-differences model (1) laid out in Section 4.3 and using a full set of controls, we estimate the potential effect of a change in the communication of incentives on the daily performance of teams in terms of the total amount of lettuce harvested per day and team (thereby, conditioning out potentially systematic variation of controls). The treatment effect is the coefficient on a binary indicator that is equal to 1 for treated teams during the treatment period, and zero otherwise. Indicator variables for treated teams and the treatment period are absorbed by team and day indicators. The results are displayed in columns (1) and (2) of Table 2. Following the discussion at the end of Section 4.2, in column (1) the month of August is defined as the treatment period, while in column (2) the remainder of the harvest season after August 1 is defined as the treatment period.

The difference-in-differences approach reveals a statistically significant, positive treatment effect in the sense that the change in the communication of the piece rate increases daily performance in the quantity domain by 0.125–0.135 standard deviations (depending on the definition of the treatment period). This is very similar to the effect of 0.143 obtained from comparing the raw means and thus also translates into an economically sizable effect of 3.5%–3.8% relative to the unconditional mean.

The estimated coefficient on the piece rate deserves a comment. In particular, note that the negative effect of the piece rate on output should not be surprising. Similar to Shearer (2004), in the present setting the firm sets a higher piece rate when harvesting conditions are more difficult (and hence, output will be relatively low) to ensure that workers obtain an average hourly wage above the legal minimum. Hence, the piece rate does not only serve an incentive purpose but is also adjusted to the pertaining harvesting conditions. While weather conditions are likely to be taken up by day fixed effects, there still remain team-day specific factors (e.g., size and condition of the respective field, or size, maturity, and condition of the crop on a given field), which are not taken up by controls.[21] Alternatively, one might exclude the piece rate from the set of controls to avoid potential problems of "bad controls" that introduce an endogeneity problem, at the risk of omitting a relevant variable in the specification. However, it turns out that the results regarding the treatment effects are unaffected by the inclusion of the piece rate in the specification.[22]

5.1.2. Quality.

Next, we investigate the effect of the treatment on performance in terms of quality. Figure 1(b) displays the respective raw means of quality (malus) points by treatment status and treatment

Table 2. Main Results

Dependent variable:	*Pieces harvested*		*Quality (malus) points*		*Manager daily compensation*	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment period (Aug. 1–31) × Treated team	0.135** (0.029) [0.03;0.24]		0.243* (0.081) [0.01;0.51]		0.178*** (0.002) [0.03;0.32]	
Treatment period (after Aug. 1) × Treated team		0.125** (0.014) [0.05;0.19]		0.095 (0.465) [−0.16;0.34]		0.155** (0.019) [−0.08;0.41]
Piece rate	−0.195*** (0.001)	−0.193*** (0.001)	0.027 (0.553)	0.026 (0.580)	−0.008 (0.749)	−0.006 (0.817)
Manager dummies	Yes	Yes	Yes	Yes	Yes	Yes
Day dummies	Yes	Yes	Yes	Yes	Yes	Yes
Team controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,182	1,182	1,182	1,182	1,182	1,182
Adjusted R^2	0.890	0.890	0.446	0.444	0.795	0.795

Notes. The table reports OLS estimates. The *p*-values (in parentheses) are based on robust standard errors that allow for clustering on the day level. In addition, in square brackets, we report 95% confidence interval that are based on standard errors that allow for clustering on the day and team level, where the team dimension is bootstrapped (with 500 repetitions). Observations are on the team-day level, where, for confidentiality reasons, all variables (except for dummy variables and fractions) are standardized to a mean of zero and a standard deviation of one on the entire sample of 1,182 observations. For any given team and day, "team controls" include (i) workers' average age (in years), (ii) workers' average total tenure (in number of seasons), (iii) workers' average tenure in the current season (in days), (iv) team size (in number of workers), (v) total work hours, and (vi) a dummy variable indicating whether on the given day the team had a second shift. Due to arrival and departure of (temporary) workers, these team controls vary on the team-day level.

*, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

period. In the control period, the raw means of quality are −0.095 and 0.085 for control teams and treated teams, respectively; implying a difference of −0.18. In the treatment period of August 2008, the respective means are −0.124 and 0.266; implying a difference of −0.39. Hence, the resulting treatment effect on the treated corresponds to 0.21 standard deviations (which corresponds to an increase in quality (malus) points by 4.7%). That is, the comparison of the raw means suggests that, while the quality performance of control teams does hardly vary across the control period and the treatment period of August, the number of quality (malus) points of treated teams goes up (and hence quality goes down) in August.

To more precisely determine the treatment effect on quality, we estimate the difference-in-differences model (1) with a full set of controls. Results are displayed in columns (3) and (4) of Table 2, where we use the same specifications as for quantity. Here, the dependent variable is quality (malus) points; and recall that a higher number of points corresponds to lower quality. As suggested by the comparison of raw means, we observe an increase in quality (malus) points by 0.1–0.24 standard deviations (which corresponds to 2.2%–5.4% lower quality). However, the treatment effect is still fairly imprecisely measured. Overall, the performance of the empirical model is weaker for quality than for quantity, which is also suggested by the comparably low R^2. This might partly be driven by the discrete (and coarse) nature of quality (malus) points as the firm's quality measure and by noise in their measurement (due to random sampling by the firm's quality control staff).

Nonetheless, when interpreting the treatment effects qualitatively, there is an indication that a slight change in the communication of incentives leads to a higher priority for quantity at the (potential) cost of quality. To investigate whether the treatment had any material effects on team managers, in a next step, we look at how it affected managers' pay.

5.1.3. Manager Daily Compensation. Ceteris paribus, a manager's pay is increasing in both quantity and quality, which suggests that the treatment had countervailing effects on manager daily compensation. We first illustrate the effect of the treatment on manager daily compensation in Figure 1(c), which displays the raw means for the two groups and periods. The comparison of raw means during the control period reveals a difference between control teams and treated teams of 0.365 (0.168 versus −0.197). In August, this difference is 0.212 (0.075 versus −0.137), which implies a positive treatment effect on the treated of 0.153 standard deviations (which corresponds to an increase in daily compensation of about 4.1%). Estimation results for the effect of the treatment on manager daily compensation using the full set of controls are reported in columns (5) and (6) of Table 2. We find a statistically significant, positive treatment effect of 0.16–0.18 standard deviations, which, again, is similar to what is suggested by the comparison of raw means. Economically, the effect corresponds to a rise in daily pay of 4.3%–4.8%. This finding seems to indicate that the treatment not only affected harvesting (in terms of quantity and quality) but also allowed managers to increase their pay. This issue will be studied in more detail in Section 5.3.

5.1.4. Economic Significance of Effects. To summarize, we find that, as a response to the treatment, quantity goes up by 3.5%–3.8% relative to the (unconditional) mean and manager daily compensation goes up by 4.3%–4.8% relative to the (unconditional) mean. Depending on the definition of the treatment period, the reduction in quality is marginally significant or insignificant, at the order of 2.2%–5.4% relative to the (unconditional) mean. To put these effects into perspective, note that to achieve a comparable increase in quantity, the firm would have to add one additional worker to each team—having to bear the cost of this additional worker.[23] Hence, the estimated effects appear to be economically sizable.

5.2. Robustness

In this section we perform various checks that document that our main results as presented in Table 2 are robust. All respective regression tables are relegated to the online appendix.

5.2.1. Alternative Definitions of Treatment and Control Periods. As discussed in Section 4.2, there are different ways of thinking about the treatment and control periods. Technically, the experiment has an "ABA withdrawal design," but due to the nature of the intervention, it is not fully clear how to think of the period after August 31 (i.e., after withdrawal of the treatment) in the empirical analysis. In Section 5.1 we reported results for a specification reflecting the "ABA" design as well as for the opposite extreme of what might be called an "ABB" specification.

As discussed in Section 4.2, as our treatment concerns a variation in the communication of incentives, one could argue that the remainder of the harvest season after August 31 is not a control period but should be viewed as part of the treatment period: To get around this subtlety, we demonstrate that our results still hold when we restrict attention to the sample up to August 31 (see Table R.3 in the online appendix).

Alternatively, one might estimate a more flexible specification allowing for a separate coefficient for the period after treatment by including two separate interaction coefficients of treated teams with the effect of the treatment period (August) and the posttreatment period (after August). The respective results for such an "ABC" specification are contained in Table R.4 in the online appendix. The effects are smaller than during the treatment period of August, but the main results are unaffected. In particular, as would be expected, the estimates of the "ABA" and "ABB" specifications reported in Table 2 deliver smaller treatment effects than the results from the "AB" or "ABC" specifications. The reason is that, in the "ABA" specification, part of the treatment effect (that persists after August) is attributed to the control period, or, in the "ABB" specification, part of the control period (after the withdrawal of the treatment) is attributed to the treatment period.

5.2.2. Persistence. Since Gneezy and List (2006), the persistence of effects has been one focus in the discussion of field experiments. The persistence of our treatment effect is indirectly shown by the results for the different specifications of the treatment period, in particular in the results that allow for a distinct effect during August and during the period after the treatment in August shown in Table R.4 in the online appendix. In particular, the results suggest a persistent effect that becomes weaker once the treatment is withdrawn.[24]

5.2.3. Manager-Specific Linear Time Trends. In our main specification, we include full sets of manager dummies and day dummies. In Table R.5 in the online appendix, we report regressions where we consider manager-specific linear time trends in addition to manager and day dummies.[25] This specification is, therefore, even more flexible than the baseline specification. Manager-specific time trends capture potentially different dynamics of managerial performance and thus rule out concerns about underlying dynamics driving the results. Panel A (panel B) of Table R.5 presents the results from specifications that include manager dummies and manager-specific linear time trends (manager dummies, manager-specific linear time trends, and day dummies); in both cases the estimates are similar to those obtained with the baseline specification.

5.2.4. Alternative Formats of Variables. In Table R.6 in the online appendix, we report regressions where the nonstandardized variables (except for dummy variables) have been log-transformed, and coefficients can be interpreted as elasticities. The results are qualitatively identical and even quantitatively very close to the results obtained with standardized variables. In unreported regressions, we have verified that the results are also robust when using the nonstandardized variables directly.

5.2.5. Alternative Clustering of Error Terms. The results in Section 5 are based on clustering of error terms on the day level, or on two-way clustering on day and team. Clustering by day is appropriate as it is likely that there are systematic day-to-day variations in the data due to, for example, weather conditions, such that outcomes for managers co-move due to unobserved factors that are not picked up by day fixed effects. This suggests that harvest conditions (such as weather) are correlated across teams on a given day. Clustering on the team level, or on the day-team level, yields identical point estimates but slightly different standard errors. The problem with clustering on the team level is the relatively small number of clusters, which is why we apply a bootstrap procedure for the team clusters. The results in Table 2 show that we obtain similar results, but somewhat lower significance levels in the case of day-team clustering. The same holds when clustering on team only. Overall, the results suggest, however, that systematic correlation in errors is unlikely to deliver spurious inference.[26]

5.2.6. Generalized Least Squares (GLS) Estimates. To document that our results are not driven by the OLS specification, we also produced GLS estimates that allow for team-specific AR(1) disturbances and heteroscedasticity across teams. Results show that the GLS results parallel the OLS findings for quantity, quality, and manager daily compensation: quantity increases significantly in response to the treatment, quality decreases (but the effect is less precisely estimated), while manager daily compensation increases significantly.[27]

5.2.7. Placebo Tests. To check whether the treatment picks up some spurious effects, we perform several placebo tests. First, we consider placebo treatment periods and counterfactually define "July 1–31" as treatment period relative to the entire observation period or the period before August 1. As revealed by panel A of Table R.8 in the online appendix, neither of the treatment effects is significant at a conventional level for either of the two placebo treatment periods. Second, as a further placebo test, instead of considering the actually treated teams, we randomly draw five teams and proceed as if they had been treated. Again, it is reassuring that panel B of Table R.8 shows that neither of the treatment effects is significant given a placebo selection of treated teams.

5.3. Behavioral Responses to the Treatment

So far, we have documented that there are systematic effects of the intervention on the main outcome variables, i.e., a higher quantity harvested and higher manager daily compensation (with imprecisely measured adverse effects on quality). In this section we investigate which behavioral responses to the treatment might have led to these effects.

5.3.1. Task Assignment by the Manager. Within his team, the respective manager has all relevant decision rights on the field. If a manager intends to harvest a larger quantity, he has only a limited number of ways to achieve this. First, he could communicate to workers that he deems a lower quality threshold acceptable (thereby pushing quantity). However, while suggestive, we only find noisy evidence for effects of the treatment on quality (see columns (3) and (4) of Table 2). Second, he could make individual cutters (i.e., the workers who do the actual cutting and hence most directly influence the quantity harvested) work harder (e.g., by raising the speed of the harvest machine). We can look into this because the quantity performance of cutters is measured at the individual level (by counting the plastic bags, in which they put the lettuce heads). However, unreported regressions show that there is no treatment effect on the pieces harvested by individual cutters. Finally, as he has the authority to decide on task allocation within his team, if a manager deems a higher quantity desirable, he could assign a larger fraction of workers to do the actual cutting (and a lower fraction to the packing and processing of the harvested crop). To investigate this latter channel, Table 3 reports estimation results where we regress the fraction of workers in the team that act as cutters on the treatment variables and the same full set of controls as before. Columns (1) and (2) of Table 3 (which only differ in the definition of the treatment period) show that as a response to the treatment, managers indeed seem to refocus attention on quantity by assigning a significantly larger fraction of their workforce to the role of cutter—the most physically demanding task—than they otherwise would have done. Economically, the coefficients in columns (1) and (2) imply that the treatment increased the fraction of cutters by roughly one percentage point. A back-of-the-envelope calculation on the basis of the nonstandardized data indicates that this increase on average corresponds to one third of a worker additionally being assigned to be a cutter, which, given the average quantity performance per cutter, approximately accounts for the entire treatment effect on quantity.

Table 3. Task Assignment by the Manager

Dependent variable:	*Fraction of cutters in the team*	
	(1)	(2)
Treatment period (*Aug.* 1–31) × *Treated team*	0.009*** (0.002) [0.002; 0.017]	
Treatment period (*after Aug.* 1) × *Treated team*		0.010*** (0.001) [0; 0.020]
Piece rate	−0.002 (0.130)	−0.001 (0.171)
Manager dummies	Yes	Yes
Day dummies	Yes	Yes
Team controls	Yes	Yes
Observations	1,182	1,182
Adjusted R^2	0.484	0.487

Note. The note below Table 2 applies.

5.3.2. Responsiveness of the Performance to Quantity Incentives. The experimental intervention led to a stronger emphasis of managers on quantity output. As discussed in Section 3.2, the firm's headquarters sets the piece rate with the dual purpose of providing quantity incentives and adjusting for varying harvesting conditions, and there is variation in the piece rate at the team-day level. Hence, a change in the communication of quantity incentives might not only lead to higher quantity per se, but it might also be that, as a result of the intervention, managers more finely adjust their behavior to variations in the piece rate (which, according to columns (1) and (2) of Table 2, significantly affects quantity). If this channel was indeed in effect, the treatment should lead to quantity output being more closely correlated with the piece rate.

Table 4 presents results that test this hypothesis by estimating the baseline specification of Table 2 where, in addition, we include interaction terms between the treatment condition (*Treatment period* × *Treated team*) and the piece rate. It turns out that the interaction terms in columns (1) and (2) are positive (and of similar size). This means that the treatment effect on quantity is stronger (weaker) if piece rates are higher (lower), i.e., quantity incentives are stronger (weaker). However, only in column (2) the interaction effect is significant at the 5% level. These findings might cautiously be interpreted as indicative of more fine-tuning of quantity performance to varying incentives in response to the treatment.

Table 4. Responsiveness of the Quantity Performance to the Piece Rate

Dependent variable:	*Pieces harvested*	
	(1)	(2)
Treatment period (*Aug.* 1–31) × *Treated team*	0.162** (0.024) [0.026; 0.298]	
Treatment period (*Aug.* 1–31) × *Treated team* × *Piece rate*	0.104 (0.312) [−0.067; 0.288]	
Treatment period (*after Aug.* 1) × *Treated team*		0.131** (0.010) [0.059; 0.203]
Treatment period (*after Aug.* 1) × *Treated team* × *Piece rate*		0.143** (0.012) [0.044; 0.246]
Piece rate	−0.201*** (0.001)	−0.243*** (0.001)
Manager dummies	Yes	Yes
Day dummies	Yes	Yes
Team controls	Yes	Yes
Observations	1,182	1,182
Adjusted R^2	0.890	0.893

Note. The note below Table 2 applies.

5.3.3. A Closer Look at Manager Daily Compensation. In a next step, we look in more detail into how the increase in manager daily compensation comes about. On the one hand, it could be that the positive effect on manager daily compensation is mechanically driven by the positive treatment effect on quantity. On the other hand, managers take all of the main (operative) decisions about the relevant harvest parameters (e.g., the speed of the harvest machine or the allocation of workers to tasks), which, in principle, allows them to fine-tune their behavior. This leads to the question whether the treatment might potentially have led managers to respond better overall to incentives. Table 5 aims to shed light on this. There, we report estimations where manager daily compensation is regressed on quantity, quality, the treatment condition (*Treatment period* × *Treated team*), and the same set of control variables as before. Column (1) of Table 5 confirms that, as suggested by the design of the incentive system, manager daily compensation is increasing in pieces harvested and decreasing in quality (malus) points. In columns (2) and (3), we additionally include the treatment condition for the two definitions of the treatment period under consideration. Reassuringly, the coefficients on pieces harvested and quality (malus) points are very stable across columns (1)–(3). The positive treatment effects in columns (2) and (3) indicate that, as a result of the intervention, managers were able to raise their pay beyond the direct effects of quantity and quality.

5.3.4. Dynamic Structure of the Treatment Effects. Finally, we study how the treatment effects evolve over time. To investigate this, we split the treatment period of August 2008 into two-week subperiods to see whether we find differential effects. Table 6 presents results for this specification (where the same full set of controls as in Table 2 is employed). In the second half of August, the treatment effects on quantity (column (1)) and manager daily compensation (column (3)) are positive and significant, while in the first half of August they are smaller and insignificant. The treatment effect on quality (column (2)) is stable across the subperiods, but depending on the specification of the error structure weakly significant in the second half of August only. Similarly, with respect to the fraction of cutters in the team, there is a positive treatment effect in both subperiods.[28] Taken together, these findings suggest that the treatment effects did not fully set in right away but needed some time to build up.

6. Potential Mechanisms

In Section 5 we established the treatment effects on the main outcome variables, and we have explored the behavioral responses that the intervention appears to have triggered. In the present section, we consider various potential mechanisms that might explain why the treatment had the observed effects.

Table 5. A Closer Look at Manager Daily Compensation

Dependent variable:	*Manager daily compensation*		
	(1)	(2)	(3)
Treatment period (Aug. 1–31) × Treated team		0.137**	
		(0.009)	
		[−0.013; 0.278]	
Treatment period (after Aug. 1) × Treated team			0.108*
			(0.080)
			[−0.143; 0.360]
Pieces harvested	0.429***	0.423***	0.423***
	(0.001)	(0.001)	(0.001)
Quality (malus) points	−0.066***	−0.067***	−0.066***
	(0.002)	(0.002)	(0.002)
Piece rate	0.074**	0.076**	0.077**
	(0.042)	(0.040)	(0.035)
Manager dummies	Yes	Yes	Yes
Day dummies	Yes	Yes	Yes
Team controls	Yes	Yes	Yes
Observations	1,182	1,182	1,182
Adjusted R^2	0.817	0.817	0.817

Note. The note below Table 2 applies.

6.1. Managers Infer Additional Information

One reason why treated managers might have changed their behavior could, in principle, be that the treatment conveyed additional information to them; not in terms of hard, material information on the incentive system but in terms of the firm's attitude toward their quantity performance.

On the one hand, if treated managers took the intervention as a signal that the firm was not satisfied with their pretreatment quantity performance (relative to the control teams), this might have led to the observed effects.[29] However, three aspects make it implausible that treated managers held such a belief. First, in Table 7, we report on regression results that document that in the pretreatment month treated teams are not significantly different from control teams in terms of either quantity, quality, or manager daily compensation once one controls for observables. Hence, there is no evidence that treated managers were "bad" managers relative to the control group.[30] Second, in various discussions with the firm's management, we were assured that the firm has a rather direct way of communicating concerns. If the firm would have been dissatisfied, managers would have expected a less subtle approach. Third, extensive discussions with the firm's management suggest that in the context of lettuce harvesting, incentives are essentially exclusively monetary, rendering explanations based on beliefs about

Table 6. Dynamic Structure of the Treatment Effects

Dependent variable:	*Pieces harvested* (1)	*Quality (malus) points* (2)	*Manager daily compensation* (3)	*Fraction of cutters in the team* (4)
Treatment period (Aug. 1–15) × Treated team	0.018	0.253	0.106*	0.009**
	(0.803)	(0.285)	(0.100)	(0.041)
	[−0.111; 0.157]	[−0.052; 0.585]	[−0.014; 0.243]	[−0.002; 0.022]
Treatment period (Aug. 16–31) × Treated team	0.236***	0.235*	0.240***	0.010***
	(0.002)	(0.071)	(0.001)	(0.006)
	[0.083; 0.382]	[−0.074; 0.572]	[0.034; 0.454]	[−0.002; 0.020]
Piece rate	−0.196***	0.027	−0.009	−0.002
	(0.001)	(0.551)	(0.734)	(0.130)
Manager dummies	Yes	Yes	Yes	Yes
Day dummies	Yes	Yes	Yes	Yes
Team controls	Yes	Yes	Yes	Yes
Observations	1,182	1,182	1,182	1,182
Adjusted R^2	0.890	0.446	0.795	0.484

Note. The note below Table 2 applies.

Table 7. Pretreatment Differences Between Control and Treated Teams

Dependent variable:	*Pieces harvested* (1)	*Quality (malus) points* (2)	*Manager daily compensation* (3)
Treated team	0.011	0.506	−0.534
	(0.974)	(0.408)	(0.348)
Piece rate	−0.308***	−0.084	0.099
	(0.006)	(0.752)	(0.603)
Manager dummies	Yes	Yes	Yes
Day dummies	Yes	Yes	Yes
Team controls	Yes	Yes	Yes
Observations	216	216	216
Adjusted R^2	0.905	0.390	0.787

Note. The sample period is the pretreatment month of July 1–31, otherwise the note below Table 2 applies.

informal sanctions or gift-exchange and reciprocity between workers and the firm unlikely.

On the other hand, treated managers might have interpreted the intervention as the firm being, in general, dissatisfied with the overall quantity performance. However, the firm's established way of communicating directives or feedback on performance suggests that it is unlikely that managers (who had worked for the firm under the same incentive system for several years) held such a belief. Importantly, at no point in the harvest season the firm issued any communication to this effect.[31]

6.2. Hawthorne Effect

Another potential rationalization for the results could be some form of "Hawthorne Effect," i.e., it could be that there is a response simply due to the fact that there was *some* intervention. However, in the present context, various pieces of evidence suggest that such an explanation is not the most plausible in our context. First, the intervention is minimally invasive and constitutes no explicit signal or change in the way managers and workers are treated. In particular, the experiment did not come along with increased scrutiny as managers and workers were fully aware at any point that quantity and quality were measured as before, and as usual their pay was based on these measures. Second, teams do not only react to the treatment by harvesting a higher quantity per se, but they seem to display a more elaborate response to the posting of the piece rate. In particular, as Table 4 indicates, they appear to more strongly fine-tune their behavior to fluctuations in the piece rate (e.g., harvest less when the piece rate is lower). In a similar vein, as a response to the treatment, managers not only harvest a higher quantity but also seem to better optimize against the incentive scheme, thereby realizing a higher compensation (as is evident from the significant positive treatment effects in columns (2) and (3) of Table 5).

Finally, following Levitt and List (2011) evidence for Hawthorne effects is slim. Hong et al. (2013) is a rare example for a cleanly documented Hawthorne effect. However, unlike in the present case, their effect is the response to a placebo letter expressing appreciation for the worker's contribution to the firm, i.e., arguably something that might affect the worker's work attitude.

6.3. Workers Infer Additional Information

While both the respective manager and his workers had every incentive to learn the pertaining piece rate, in principle, it could be that, absent the treatment, managers failed to communicate this important piece of information to their workers at the beginning of the respective shift. If this indeed would have been the case, the intervention would also have provided workers with additional information, which could be yet another mechanism causing the observed treatment effects. However, again, various aspects make it unlikely that this is a comprehensive explanation for our findings. First, given that both managers' and workers' pay is heavily incentive-based, we deem such a lack of communication as very unlikely.[32] Second, the evidence on task allocation in Table 3 indicates that as a response to the treatment managers change their behavior (i.e., it is also managers who seem to act differently in response to the treatment).[33] Finally, one might suspect that workers holding additional information would only have effects if they were able to affect what happens on the field (i.e., influence the manager's decision making). In turn, if they were indeed able to do so, one would expect them to also benefit from any implemented changes. However, additional unreported regressions display no treatment effect on worker daily compensation (or on the compensation of the subset of cutters within a team).

6.4. Higher Salience of (Quantity) Incentives to Managers

Based on the discussions in Sections 6.1, 6.2, and 6.3, we argue that the mechanisms discussed so far cannot fully explain the available evidence, suggesting that there is another (additional) mechanism at work.

Given the evidence from other domains, such as personal finance or consumer choices (see Section 2 on the related literature), and given the many demands on a manager's time and attention in the present multitasking context (see Section 4.2), the treatment might have made (quantity) incentives and the role of quantity in the firm's priorities more salient to managers, thereby affecting behavior. Such a salience-based explanation (as laid out in the following) would be consistent with all of the available evidence: The posting of the piece rate during the treatment period seems to have led managers to refocus on the specifics of the incentive system, to adapt their behavior (e.g., choose a different allocation of tasks within the team and more finely tune behavior toward variations in the piece rate), and to achieve a higher compensation.[34] As the posting of the piece rate was repeated on every day of August, it is not implausible that higher salience of the incentives gradually built up, which would be consistent with the dynamic structure of the documented treatment effects.[35]

This is not to claim that the alternative mechanisms do not play a role in general. However, they seem to be either inconsistent with some of our findings or implausible in the current setting. Hence, our experiment indicates that, even in the context of incentive provision for experienced managers, salience might play an important role.

7. Discussion

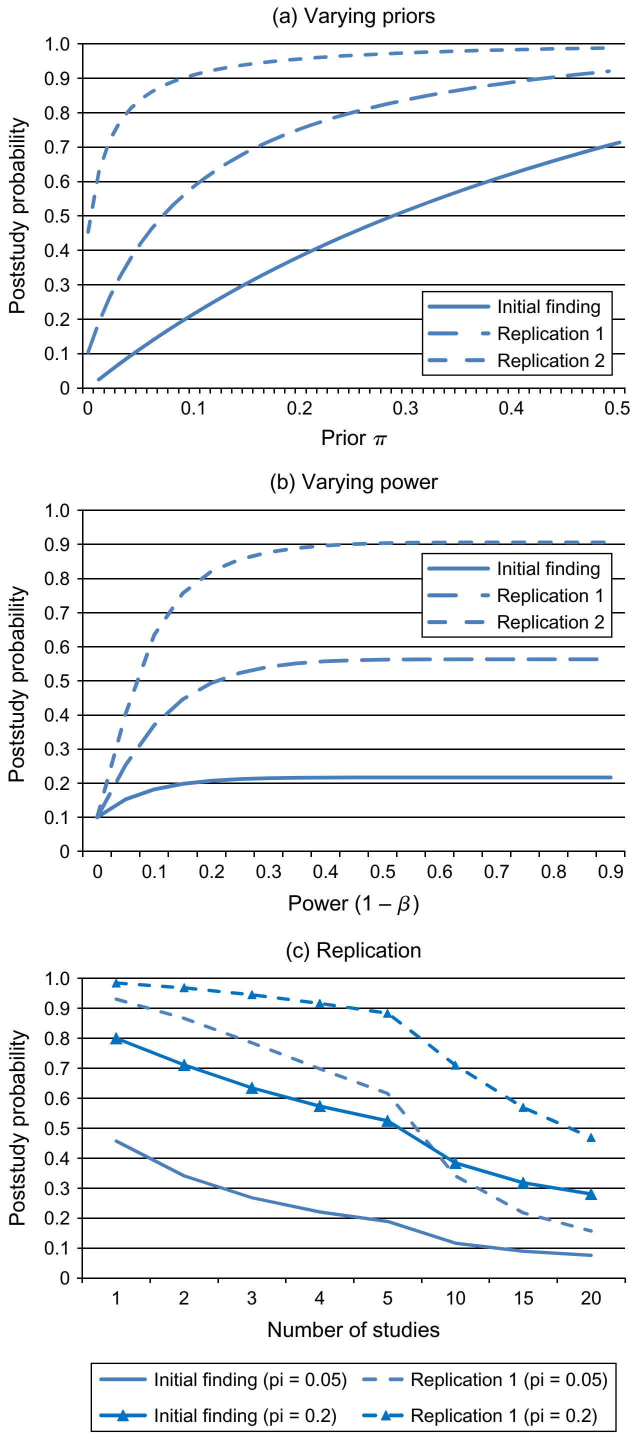
This paper reports evidence from a randomized intervention in a real-world team production setting where experienced agents work under a sophisticated incentive system in a high-stakes environment.

Our findings indicate that a minimally invasive change (a reminder) in the way an important component of the incentive system, the piece rate, is communicated (while keeping the material incentive system unchanged) has statistically significant and robust effects on performance. In particular, we find that the change in the communication of the piece rate component of the incentive system increases output (quantity) and manager compensation, while having an adverse (though less precisely measured) effect on quality. These economically sizable effects—for example, the treatment effect on quantity corresponds to what would be achieved by adding one additional worker to the team—seem to be moderated via a changed assignment of tasks by the manager within the team and an increased responsiveness of output to variations in the piece rate. In light of our discussion of various alternative interpretations of our findings, we argue that a salience-based mechanism appears as the most plausible explanation for the entire set of empirical findings: By repeating relevant information that was already available to managers, the intervention may have increased the managers' awareness of the particular incentives on a given day and the information contained in its day-to-day adjustments.

At first sight, the result of treatment effects where, due to no change in material incentives, standard theory would predict no reaction, appears surprising. This ties into a more general, emerging discussion in economics about the appropriate interpretation of empirical findings. The main concern in this context is that empirical results such as those presented here, despite being statistically significant and robust across a variety of specifications, might not reveal a "true" association but instead reflect a "false positive." As pointed out by Maniadis et al. (2014, 2015), it is inherent to the mechanics of statistical inference that the probability that a research finding actually reveals a true effect (i.e., the *poststudy probability*) does not only depend on statistical significance, but, importantly, also on the prior one assigns to finding such an effect, as well as statistical power. The prior probability of an effect of our change in the communication of incentives might arguably be small. Moreover, the findings are based on a moderate sample of 10 independent teams that were followed over an entire harvest season (about 150 days), half of which experienced the treatment. The poststudy probability of the effect might thus be moderate, raising concerns about the overall informativeness of the results. On the other hand, if the effects that we document in the present field experiment indeed reflected true associations, the findings would have important implications for firms as they would indicate that details of the communication of incentives are an important aspect even for experienced workers employed under an elaborate incentive system. In the following, we elaborate on this important point in more detail and discuss the potential gains from replications of our field experiment.[36]

Based on Maniadis et al. (2014), one can, in fact, evaluate the relevance of our findings in terms of how they change the priors on the phenomenon. As stated in the Introduction, there is only a moderate number of studies that report effects of variations in the communication of relevant aspects of decision problems (where these existing studies mostly focus on contexts in personal finance or on consumption choices; we are not aware of any prior study on the effect of communication of incentives in an employment context). Given that these are likely to reflect the number of studies that found significant results out of a potentially larger (and unknown) number of studies that have investigated similar settings, one might ask by how much the present results increase the probability that finding a significant effect indeed reflects a "true" effect. Figure 2 depicts this poststudy probability in different ways.[37] Panel (a) depicts the poststudy probability for varying priors of a true effect. The graph illustrates that a first finding of an effect raises the poststudy probability considerably but to a varying extent, depending on the prior. However, even only one or two successful replications lead to a substantial increase in the poststudy probability of a true effect. For instance, when the prior of a true effect is 0.2, the poststudy probability after an initial finding (from a total of 10 studies) is 0.38, i.e., almost doubled. A replication of the finding (in a second of the 10 studies conducted on the subject) raises the poststudy probability to 0.74, and a second replication raises it to 0.95. For a prior of 0.05, the value of replication is even more striking. The poststudy probability of 0.12 from a first finding rises to 0.38 after a first replication, and to 0.81 after a second

Figure 2. (Color online) Poststudy Probability



Notes. Poststudy probability as function of prior, power, and the number of studies on the subject. See the main body of the paper and Endnote 37 for details. In the plots, the significance level is set to $\alpha = 0.05$. In panel (a), the other parameters are $n = 10$ and $(1-\beta) = 0.8$; in panel (b), the parameters are $n = 10$, and $\pi = 0.2$; and in panel (c), the statistical power is assumed as $(1-\beta) = 0.8$.

replication. Panel (b) illustrates that these observations are essentially unaffected by statistical power, unless power is below 0.1. Panel (c) illustrates the gain in confidence that can be obtained by a successful replication of a result depending on the number of studies on the subject. The gain of replication is particularly large if relatively few studies have investigated a similar issue, which is very likely the case in the context of slight changes in the communication of incentives in the workplace.

Hence, under the assumption that the overall number of studies on the subject (which corresponds to the experiments conducted, not the published results, and hence is unobserved) is small, our finding of a significant effect likely implies a substantial revision of the poststudy probability of a true effect. This is particularly the case in the context of a workplace setting with experienced workers and no changes in an established incentive system, given that our study is, at least to our knowledge, the first that investigates the effect of this kind of intervention in such a workplace environment. Figure 2 also suggests that additional replication studies can deliver further substantial improvements in the poststudy probability, particularly in light of the low prior and the moderate power in the context of our study. Hence, we view the evidence presented in this paper as suggestive of a potentially important effect; calling for further investigation by ways of replication in comparable settings. Certainly, more work is needed to pin down the effect of variations in the communication of incentives and the relative importance of different potential explanations.

In this respect, recent theories that model the effects of limited attention or salience on behavior (see e.g., Bordalo et al. 2012 or Koszegi and Szeidl 2013) might be useful to design follow-up experiments. Broadly speaking, in these models, the decision maker's attention is drawn to payoffs that markedly "stand out," and decisions are accordingly tilted toward these salient payoffs. Our preferred interpretation of the empirical findings is consistent with these theories. The lump sum prize of winning the quality tournament "stands out" and hence might have directed excess attention to the quality dimension, while the treatment might have re-focused attention to reflect both dimensions of the multitasking problem. By extending these recent theories to problems of incentive design, it should be possible to derive more specific predictions relevant to firm contexts, which could be further explored and tested in appropriately designed (field) experiments. However, the precise identification of the mechanisms through which these effects operate might be difficult within minimally invasive field experiments and will probably require more invasive interventions than the natural field experiment presented in this paper.

Acknowledgments

The authors thank Josh Angrist, Oriana Bandiera, Iwan Barankay, Jordi Blanes-i-Vidal, Stefano DellaVigna, Martin Huber, Nicola Lacetera, Imran Rasul, Matthias Schündeln, Robert Ulbricht, and Fabian Waldinger, as well as seminar participants at Chicago, European School of Management and Technology (ESMT) Berlin, Innsbruck, Institute for the Study of Labor (IZA) Bonn, Kellogg, Massachusetts Institute of Technology, Toulouse, Sydney, and various conferences for helpful comments and suggestions.

Endnotes

[1] See also, for example, Paarsch and Shearer (2000), Shearer (2004), Bandiera et al. (2005), and Bellemare and Shearer (2011). See Prendergast (1999) for a survey of the earlier literature. Charness and Kuhn (2010) provide an overview of the (lab and field) experimental evidence on behavioral aspects and unintended consequences of incentive provision.

[2] Our study is a natural field experiment as it occurs in the environment where the subjects naturally undertake these tasks, but where subjects do not know that they are participants in an experiment; see, Harrison and List (2004). Following Al-Ubaydli and List (2012, 2013), natural field experiments combine two important elements of the experimental method and of naturally occurring data: randomization and realism. In addition, as subjects are randomly placed into treatment and control—and randomization balances the unobservables—they tackle the typical selection effects discussed in the literature. Hence, we can interpret the parameter estimate of the treatment effect as the causal effect of the treatment intervention.

[3] The pay of the (temporary) harvest workers is entirely incentive based. The amount of money they earn while working for the firm constitutes a substantial fraction of their annual income.

[4] As will be discussed in more detail in Section 5.1, the difference in findings for quantity and quality might be due to the fact that quality is less precisely measured. We also discuss these findings in the light of the recent empirical literature on multitasking effects; see Section 2.

[5] See Kluger and Denisi (1996) for a survey of the psychology literature on the effects of information interventions on performance.

[6] For a recent contribution that uses "empty messages," see Cohn et al. (2016). They, however, use an empty message as control while the treatment message contains material information about a change in (gift-exchange based) incentives.

[7] New workers typically work on a fixed daily wage for one to two days (while practicing their task) and switch to incentive pay thereafter.

[8] Although there are considerable differences in the production technology (and the respective product), the composition of the workforce is comparable to the one in Bandiera et al. (2005).

[9] For example, a Polish worker on average earns more than 40% of the Polish average annual salary during a typical spell at the firm.

[10] In practice it is common that piece rates fulfill such purposes; see, for example, related applications in Paarsch and Shearer (2000) and Shearer (2004).

[11] The total number of work hours includes the hours worked by the manager, and (in addition to additional pay components, which are discussed in Section 3.2.4) the manager receives the resulting average hourly pay from quantity incentives. For cutters (whose quantity performance is observed at the individual level), there are additional performance-dependent adjustments.

[12] We did not implement an "active control" condition, such as an empty sheet or a note urging managers to focus on efficiency, for several reasons. First, in the design stage, the firm clearly communicated that the intervention would have to be minimally invasive, minimizing the additional work load of the quality control teams. Second, by adding a variation also for control teams, the interpretation of the effect would be unclear, obstructing the difference-in-differences comparison between treatment and control.

[13] Note that teams operate on different fields, dispersed over a 10–15 km radius, so that in general they are too far apart to communicate with each other during a given shift. This implies that spill-overs from direct observations can be ruled out. Also, recall that there is no rotation of workers across teams. When workers arrive at the farm, they are assigned based on current vacancies to a team, and it is firm policy that they stay with this team. Finally, none of the team managers experienced both treatment conditions. At the same time, managers know each other well (since they regularly meet in weekly meetings held in the firm's headquarters), and communication among teams about incentives (or the intervention itself) after or before work cannot be ruled out. In principle, through such communication the treatment might even affect non-treated teams. However, managers typically do not see each other every day, because the shifts begin and end at different times, and fields are spread out in the region. Hence, the most likely occasion for such conversations to take place was during the regular meetings with the headquarters staff. As it is not in the interest of control teams to slack off, but, if anything, to work as if treated to preserve their position in the quality tournament, communication during these meetings should bias any treatment effect toward zero. Consequently, the results presented in Section 5 might be viewed as conservative estimates of the true treatment effects. Moreover, the headquarters staff gave us no indication whatsoever that such conversations had taken place.

[14] On less than 5% of harvest days, a given team had already completed an earlier shift on a different field on the same day (where such second shifts tend to be short). The analysis in Sections 5 and 6 is based on a sample that drops these 53 (second) observations of a given team on a given day, and controls for this fact with a dummy variable indicating whether the team worked an additional second shift on the respective day. When calculating manager daily compensation (which will be one of our main outcome variables), we also exclude earnings from potential second shifts. Hence, we base our analysis on variation on the team-day level. Including the second shifts in the estimation sample (or, alternatively, dropping all teams with multiple shifts on a given day) delivers virtually identical results, which are available upon request.

[15] As discussed in Section 5.2, this standardization does not affect results.

[16] Moreover, we control for break time in all our regressions.

[17] Recall that, for any given team, these team controls vary over the course of the harvest season as there is continuous arrival and departure of seasonal workers.

[18] Note that it is the firm, and not the manager, who has authority to decide on these variables. Unreported regressions confirm that the treatment did not have a systematic and significant impact on the length of shifts or breaks (in hours). Finding an effect here might have indicated that (beyond the firm's directives) managers had leeway with respect to total work hours (e.g., to increase output if this seemed profitable). Details are available upon request.

[19] Also recall that the standardization is on the entire sample of 1,182 observations, and hence in any subsample the mean is not necessarily zero.

[20] Note that while Figure 1 might suggest pretreatment differences between treated and control teams, this only applies to the raw data. Once one controls for observables, there are no discernible pretreatment differences in quantity, quality, or manager daily compensation between treated and control teams. This is discussed in more detail in Section 6 (in particular, see Table 7).

[21] Relating the piece rate to the one-day-ahead forecast of the amount of rainfall (in liters per square meter) and the daily maximum temperature (in degrees centigrade) for the respective harvest day in the harvesting area reveals a positive correlation between the forecasted precipitation and the piece rate (pairwise correlation 0.14, p-value $<$ 0.01) and a negative correlation between the forecasted maximum

temperature and the piece rate (pairwise correlation before September 1 is −0.18, p-value < 0.01, and −0.07, p-value < 0.02 over the entire season). However, note that the effect of current weather on both current and future harvesting conditions is fairly intricate, and hence only certain combinations of weather conditions will affect harvesting conditions negatively.

[22] Table R.1 in the online appendix contains the corresponding estimation results.

[23] This back-of-the-envelope calculation is based upon the estimated effect of team size on pieces harvested. Details are available upon request.

[24] In the empirical literature on salience, the persistence of effects has not been treated extensively so far, and the existing evidence is inconclusive. Chetty et al. (2009) report that demand returned to preexperiment levels after the intervention ended and interpret this as suggestive for the absence of persistent effects. Stango and Zinman (2014) document cumulative and persistent, but not permanent, effects of taking part in surveys that contain questions related to overdraft fees on actually incurring overdraft fees.

[25] Additionally allowing for manager-specific quadratic time trends does not affect results.

[26] Results for team clusters are contained in the online appendix (Table R.7). Given that our specification contains both manager fixed effects and day fixed effects, clustering standard errors on team (manager) and day might be considered overly conservative. As noted by Thompson (2011), two-way clustered standard errors have less bias but exhibit more variance. Moreover, given the panel structure with large T (days) and small N (teams), clustering on teams might be restrictive (see, e.g., Angrist and Pischke 2009, Chap. 8), which is the reason for applying the bootstrap procedure used in the main specifications. Nevertheless, the results are qualitatively unaffected by the way clustering is treated.

[27] A system (SUR) estimation of quantity and quality also delivers results that do not differ qualitatively from our OLS-based findings. Details are available upon request.

[28] Table R.9 in the online appendix replicates the same analysis splitting the treatment month of August into four (weekly) subperiods. This finer-grained analysis on the time dimension leads to less precisely estimated coefficients. The findings suggest that there is a positive effect on quantity and the fraction of cutters in the team already in the second week (but not the first), but the quantity effects in weeks 3 and 4 are about twice as large. Overall, the same pattern holds for manager daily compensation. With respect to quality, there seems to be some indication of an initial effect, which, however, vanishes after week 1.

[29] We are grateful to a referee for urging us to expand on this issue.

[30] Alternatively, when comparing the residuals of quantity obtained from estimating regression model (1) with pieces harvested as the outcome variable for July 2008 (or for the entire period up to July 31) there is no significant difference between treated teams and control teams. The same applies to quality (malus) points and manager daily compensation.

[31] An auxiliary survey could, in principle, have been helpful to investigate this in more detail. However, in the context of the current study, evaluating such a change in beliefs via a survey would have been problematic. Asking directly about (the role of) quantity incentives would immediately have made them more prominent (see, e.g., the discussion of Stango and Zinman 2014, in Section 2). Moreover, given the continuous arrival and departure of (temporary) workers throughout the harvest season, any survey would have been difficult to administer and might have revealed (to both managers and workers) that an experiment was conducted. Finally, conducting a survey was considered too disruptive by the firm.

[32] Note that according to the firm's internal guide book for managers, it is the responsibility of the manager to inform all workers about the piece rate before the shift begins.

[33] Also, recall the result, discussed in Section 5.3, that there is no treatment effect on the number of pieces harvested by individual cutters.

[34] If, absent the treatment, managers had perfectly optimized against the (identical) incentive scheme, one might expect nothing to change. Importantly, recall that piece rates do not differ systematically between treated teams and control teams neither before nor during the treatment period (see the respective discussion in Section 4.1).

[35] In the context of avoiding checking overdraft fees, Stango and Zinman (2014) document that attention is increasing in the number of "reminders" (i.e., the number of financial surveys in which their subjects took part, where each of the surveys contained overdraft-related questions).

[36] In addition, List et al. (2016) have recently pointed out the need for taking the issue of "multiple hypothesis testing" into account when interpreting empirical or experimental results, and they develop testing procedures for binary treatment contrasts that account for the joint dependence structure of the test statistics in such cases. The investigation of multiple outcomes in our study (quantity, quality, and manager daily compensation) also calls for controlling the family-wise error rate (i.e., the probability of one or more false discoveries). However, the difference-in-differences regression design of our empirical analysis prevents a direct application of the procedures by List et al. (2016). Applying the multiple testing procedure by Holm (1979), which assumes a worst-case dependence structure of the test statistics (see Romano et al. 2010, for a survey), to our main results in Table 2 delivers the following p-values for the treatment effects assuming the treatment period is given by August 1–31 (after August 1): 0.058 (0.042) for pieces harvested, 0.081 (0.465) for quality (malus) points, and 0.008 (0.038) for manager daily compensation.

[37] The graphs apply the same reasoning as Maniadis et al. (2014), who build on work in medicine by Ioannidis (2005) and Moonesighe et al. (2007). The poststudy probability (PSP) of an effect that has been found to be significant at the significance level α (reflecting the probability of reporting a true association despite it being false) is given by the ratio between the probability of finding an effect that reflects a true effect over the total probability of a significant effect, $PSP = ((1-\beta)\pi)/((1-\beta)\pi + \alpha(1-\pi))$, where π is the (prior or baseline) probability of a true effect, and $(1-\beta)$ is the power of the statistical design (i.e., β is the probability of declaring the effect false despite it being true); see Maniadis et al. (2014, p. 284). The PSP after at least r statistically significant results among n studies can then be expressed as $PSP(r) = (B(1-\beta, r, n)\pi)/(B(1-\beta, r, n)\pi + B(\alpha, r, n)(1-\pi))$, where $B(1-\beta, r, n) = \sum_{i=r}^{n}\binom{n}{i}(1-\beta)^i\beta^{(n-i)}$ and $B(\alpha, r, n) = \sum_{i=r}^{n}\binom{n}{i}\alpha^i(1-\alpha)^{(n-i)}$; see Moonesighe et al. (2007, p. 219). Hence, varying r allows computing the increase in the PSP obtained through replication. Figure 2 depicts $PSP(r)$ for various constellations of parameters.

References

Al-Ubaydli O, List JA (2012) On the generalizability of experimental results in economics. Frechette G, Schotter A, eds. *Handbook of Experimental Economic Methodology*, Chap. 20, (Oxford University Press, Oxford, UK), 420–462.

Al-Ubaydli O, List JA (2013) On the generalizability of experimental results in economics: With a response to Camerer. NBER Working Paper 19666, National Bureau of Economic Research, Cambridge, MA.

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

Angrist J, Pischke J-S (2009) *Mostly Harmless Econometrics* (Princeton University Press, Princeton, NJ).

Bandiera O, Barankay I, Rasul I (2005) Social preferences and the response to incentives: Evidence from personnel data. *Quart. J. Econom.* 120(3):917–962.

Bandiera O, Barankay I, Rasul I (2013) Team incentives: Evidence from a firm level experiment. *J. Eur. Econom. Assoc.* 11(5): 1079–1114.

Barankay I (2012) Rank incentives: Evidence from a field experiment. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Bellemare C, Shearer B (2011) On the relevance and composition of gifts within the firm: Evidence from field experiments. *Internat. Econom. Rev.* 52(3):855–882.

Blanes-i-Vidal J, Nossol M (2011) Tournaments without prizes: Evidence from personnel records. *Management Sci.* 57(10): 1721–1736.

Bordalo P, Gennaioli N, Shleifer A (2012) Salience theory of choice under risk. *Quart. J. Econom.* 127(3):1243–1285.

Brown J, Hossain T, Morgan J (2010) Shrouded attributes and information suppression: Evidence from the field. *Quart. J. Econom.* 125(2):859–876.

Cameron C, Gelbach J, Miller D (2008) Bootstrap-based improvements for inference with clustered errors. *Rev. Econom. Statist.* 90(3):414–427.

Cameron C, Gelbach J, Miller D (2011) Robust inference with multiway clustering. *J. Bus. Econom. Statist.* 29(2):238–249.

Charness G, Kuhn P (2010) Lab labor: What can labor economists learn from the lab? Card D, Ashenfelter O, eds. *Handbook of Labor Economics*, 4th ed. (North-Holland, Amsterdam), 229–330.

Chetty R, Looney A, Kroft K (2009) Salience and taxation: Theory and evidence. *Amer. Econom. Rev.* 99(4):1145–1177.

Cohn A, Fehr E, Goette L (2016) Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Sci.* 61(8):1777–1794.

DellaVigna S (2009) Psychology and economics: Evidence from the field. *J. Econom. Literature* 47(2):315–372.

Dumont E, Fortin B, Jacquemet N, Shearer B (2008) Physicians multitasking and incentives: Empirical evidence from a natural experiment. *J. Health Econom.* 27(6):1436–1450.

Finkelstein A (2009) EZ-tax: Tax salience and tax rates. *Quart. J. Econom.* 124(3):969–1010.

Gneezy U, List JA (2006) Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica* 74(5):1365–1384.

Harrison GW, List JA (2004) Field experiments. *J. Econom. Literature* 42(4):1009–1055.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian J. Statist.* 6(2):65–70.

Holmstrom B, Milgrom P (1991) Multitask principal-agent analyses: Incentive contracts, asset ownership and job design. *J. Law, Econom., Organ.* 7(Special Issue):24–52.

Hong F, Hossain T, List JA, Tanaka M (2013) Testing the theory of multitasking: Evidence from a natural field experiment in Chinese factories. NBER Working Paper 19660, National Bureau of Economic Research, Cambridge, MA.

Hossain T, List JA (2012) The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Sci.* 58(12):2151–2167.

Hossain T, Morgan J (2006) . . . Plus shipping and handling: Revenue (non) equivalence in field experiments on eBay. *B.E. J. Econom. Anal. Policy: Adv. Econom. Anal. Policy* 5(2):1–27.

Ioannidis JP (2005) Why most published research findings are false. *PLoS Medicine* 2(8):696–701.

Johnson RM, Reiley DH, Munoz JC (2015) "The war for the fare": How driver compensation affects bus system performance. *Econom. Inquiry* 53(3):1401–1419.

Karlan D, McConnel M, Mullainathan S, Zinman J (2016) Getting to the top of mind: How reminders increase saving. *Management Sci.* 62(12):3393–3411.

Kluger A, Denisi A (1996) The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary information intervention theory. *Psych. Bull.* 119(2): 254–284.

Koszegi B, Szeidl A (2013) A model of focusing in economic choice. *Quart. J. Econom.* 128(1):53–107.

Lacetera N, Pope D, Sydnor J (2012) Heuristic thinking and limited attention in the car market. *Amer. Econom. Rev.* 105(5):2206–2236.

Lazear E (2000) Performance pay and productivity. *Amer. Econom. Rev.* 90(5):1346–1361.

Lee Y, Malmendier U (2011) The bidder's curse. *Amer. Econom. Rev.* 101(2):749–787.

Levitt S, List JA (2011) Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. *Amer. Econom. J.: Appl. Econom.* 3(1):224–238.

List JA, Shaikh AM, Xu Y (2016) Multiple hypothesis testing in experimental economics. NBER Working Paper 21875, National Bureau of Economic Research, Cambridge, MA.

Maniadis Z, Tufano F, List JA (2014) One swallow doesn't make a summer: New evidence on anchoring effects. *Amer. Econom. Rev.* 104(1):277–290.

Maniadis Z, Tufano F, List JA (2015) How to make experimental economics research more reproducible: Lessons from other disciplines and a new proposal. *Res. Experiment. Econom.* 18: 215–230.

Moonesighe R, Khoury MJ, Janssens ACJ (2007) Most published research findings are false—But a little replication goes a long way. *PLoS Medicine* 4(2):218–221.

Paarsch H, Shearer B (2000) Piece rates, fixed wages and incentives effects: Statistical evidence from payroll records. *Internat. Econom. Rev.* 41(1):59–92.

Pope D (2009) Reacting to rankings: Evidence from "America's best hospitals." *J. Health Econom.* 28(6):1154–1165.

Prendergast C (1999) The provision of incentives in firms. *J. Econom. Literature* 37(1):7–63.

Romano JP, Shaikh AM, Wolf M (2010) Hypothesis testing in econometrics. *Annual Rev. Econom.* 2:75–104.

Shearer B (2004) Piece rates, fixed wages and incentives: Evidence from a field experiment. *Rev. Econom. Stud.* 71(2):513–534.

Stango V, Zinman J (2014) Limited and varying consumer attention: Evidence from shocks to the salience of bank overdraft fees. *Rev. Financial Stud.* 27(4):990–1030.

Thompson S (2011) Simple formulas for standard errors that cluster by both firm and time. *J. Financial Econom.* 99(1):1–10.