Is There Folly When Worker A Is More Productive Than Worker B? Examining Heterogeneous Responses to Individual and Group Performance Pay

by

Alexander M. Kowalski

B.A. Journalism and Mass Communication
University of North Carolina at Chapel Hill, 2010

M.C.P.
University of California-Berkeley, 2015

SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN MANAGEMENT RESEARCH

at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2019

©2019 Massachusetts Institute of Technology. All rights reserved.

Signature redacted
Signature of Author:

Department of Management
May 9, 2019

Signature redacted
Certified by: ____________________________ Thomas Kochan
George Maverick Bunker Professor of Management
Thesis Supervisor

Signature redacted
Accepted by: ____________________________ Ezra Zuckerman Sivan
*Alvin J. Siteman (1948) Professor of Entrepreneurship and Strategy*
Deputy Dean
Faculty Chair, MIT Sloan PhD Program
Is There Folly When Worker A Is More Productive Than Worker B? Examining Heterogeneous Responses to Individual and Group Performance Pay

by

Alexander M. Kowalski

Submitted to Sloan School of Management on May 9, 2019 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Management Research

ABSTRACT

Although it is commonly suggested that employees of different ability levels respond in different ways to performance pay, there is little research documenting such a relationship. In this study, I examine how employee ability moderates the effectiveness of two types of incentives in a field experiment with warehouse workers at PickInc, a large U.S.-based retailer. The incentives are distinguished by their level of reward: one is tied to individual performance, the other to group performance. I find, first, that the individual incentive is superior in improving productivity and, second, that there is variation in how employees of different ability levels respond. The relationship between individual performance pay and ability is J-shaped: weak performers respond most positively, followed by top performers, followed by middle performers. Group performance pay is negatively related to ability. The results suggest that more effective performance pay systems will need to account for variability in employee response.

Thesis Supervisor: Thomas Kochan
Title: George Maverick Bunker Professor of Management
Not everyone is moved by the same motivations at work. This paper tests that insight by examining how employees of differing ability levels respond to performance pay, a common practice used by firms to motivate their workforce.

The basis for such a test is both theoretical and methodological. Scholarly work in human resources, management, and economics proposes that employees respond differently to the same incentive depending on their preferences and traits (Rynes, Gerhart, & Parks, 2005; Heywood & Jirjahn, 2006). Employee ability, the focus of this paper, is held to be a key influence. Yet evidence of how responses to performance pay vary according to ability is limited. Most work, in fact, shows that employees of different ability levels prefer different incentives; it does not show how the same employees actually respond to the incentives once they have been offered.

Meanwhile, social science methodologists have recently proposed appropriate means of analyzing differential responses to the same intervention, or what are known as “heterogeneous treatment effects” (Imbens & Wooldridge, 2009; Imai & Strauss, 2011; Xie, Brand, & Jann, 2012). These methods, however, have not been applied to the study of performance pay, where heterogeneous treatment effects are, as mentioned above, anticipated but not well documented. Indeed, most research evaluating performance pay focuses on one effect: the average outcome among all employees or participants in the study (e.g., Jenkins, Mitra, Gupta, & Shaw, 1998).

In what follows I advance on this precedent by analyzing the results of a field experiment in which workers in multiple warehouses at the same U.S. firm participated in a performance pay program intended to boost their productivity. The primary questions guiding my analysis are: 1) do workers whose past performance indicates that they have high levels of ability respond more positively to performance pay, in terms of productivity gains, than their less able co-workers, and 2) does the outcome depend on the type of performance pay they are offered? Because worker
performance is known beforehand in my study, I am able to determine whether workers of
different ability levels respond differently to the same incentive. I can also see which workers
left or joined the firm during the experiment, which allows me to distinguish between actual
responses to the incentive and sorting that may be due to worker preferences. Finally, I am able
to assess whether any heterogeneous treatment effects, if found, depend on the type of
performance pay on offer because the workers received two different incentives, one tied to
individual performance and one tied to group performance.

My results provide evidence of heterogeneous treatment effects and thus of a differential
response to performance pay based on ability, one that changes with incentive type. The workers
who are initially the most and least productive respond most positively to the individual
incentive. Only the initially least productive workers respond positively to the group incentive.

Given these findings, future research on performance pay should be targeted toward
identifying how different types of employees respond to different types of incentives rather than
toward assessing the effectiveness of incentives in general. Investigations of this sort would,
subsequently, inform a sounder and more adaptable management practice. That is, to craft more
effective performance pay systems, there is a need not only to better measure the impacts of
these systems but also to make more specific and reliable predictions about the types of systems
that are likely to motivate workers of different abilities.

THEORY
That—under the appropriate conditions (Pfeffer, 1995; Osterman, 2011)—performance pay\(^1\) can boost employee productivity is well documented (Shaw & Gupta, 2015). An important but largely overlooked detail is that the evidence in support of this conclusion, including notable contributions by Jenkins et al. (1998) and Garbers and Konradt (2014), stems almost exclusively from studies that assess the average response of those subjected to incentives rather than the variation in their responses. This is partially due to the nature of the meta-analyses that provide the bulwark of support for the productivity-enhancing power of performance pay, as they must use similar measurements to compare findings across studies. At the same time, little empirical effort has gone into examining the possibility that there is variability in how employees respond to performance pay (Gupta & Shaw, 2014). Taken as a whole, it is thus fairer to say that existing research suggests that the *average* employee responds to financial rewards by increasing output.

We do not know, therefore, so much about whether the positive relationship between performance pay and productivity is uniform across workers or whether some respond more positively than others. What evidence we do have pertains largely to gender (e.g., Gneezy, Niederle, & Rustichini, 2003; Lavy, 2013) and personality (e.g., Krenl, 1992; Young, Beckman, & Baker, 2012). Ability is given short attention as a moderator even though, as explained below, it is thought to be a major determinant of a worker’s response to performance pay.

**Differential Responses to Performance Pay**

When it comes to performance pay, there is good reason to expect that not all workers will respond alike to the same offering. Individual preferences and traits, such as one’s appetite

\(^1\) Performance pay describes several compensation systems, such as piece rates, merit pay, and profit sharing (Milkovich, Newman, & Gerhart, 2011). In this study, I focus on the form that most explicitly links compensation to performance: incentive pay. Incentive pay remunerates workers for meeting performance standards on outcomes like quantity and quality. I use this term, “performance pay,” and “incentive” interchangeably throughout the paper.
for risk, need for achievement, and sense of self-efficacy, may affect the attractiveness of a given incentive (Gerhart, Rynes, & Fulmer, 2009). This is especially so with respect to ability, which affects the cost of effort required for an individual to meet a reward threshold and is thus a key determinant of a worker’s response to incentives.

The basis for the link between ability and performance pay is straightforward. Ability lowers the cost of effort an employee must expend to increase performance (Lazear, 1986, 2000; Booth & Frank, 1999). More able employees find it less costly to work harder to achieve the levels of performance needed to receive additional income under performance pay.

This relationship manifests in two channels (Lazear, 2000). First, because the more able require smaller income gains than the less able to induce the same amount of effort, performance pay should be more powerful among the more able. Workers of different ability levels, therefore, should respond differently to the same incentive in terms of how they change their performance. This channel is called the incentive effect. Second, performance pay should be more attractive to high ability employees than flat wages, making them more likely to select into jobs offering it than the less able. When performance pay is offered, therefore, workers of different ability levels should enter and exit the organization at different rates. This channel is called the sorting effect.

Signs of a differential response to performance pay according to ability come mostly from research on sorting effects. A host of laboratory and firm-level studies reveal that the more able tend to prefer performance pay over flat (Cable & Judge, 1994; Lazear, 2000; Cadsby, Song, & Tapon, 2007; Dohmen & Falk, 2011). Research on labor markets supports this conclusion as well: in the U.S, jobs with (individual) incentive pay tend to attract more able workers (Parent, 1999; Curme & Stefanec, 2007). While theoretically sound, the notion that the
more and less able respond differently to performance pay—that there are incentive effects—is largely inferred from these findings on sorting.

Only a small set of studies actually reveals differential responses, and their conclusions are mixed. In a lab experiment, Eriksson and Villeval (2008) find that higher skill employees increase their productivity more under performance pay than less skilled employees do. On the other hand, Franceschelli et al. (2010), studying a firm’s switch to piece rates, find that low-ability workers increase their output slightly more than better performers. These studies thus affirm the existence of a differential response but offer conflicting insights.

Such limited research on this topic is shortsighted for both researchers and practitioners. For one thing, a focus only on average outcomes can obscure variation in responses that would be informative to those trying to design optimal policies (Imbens & Wooldridge, 2009; Imai & Strauss, 2011). What’s more, conclusions about the effectiveness of a given program can vary depending on the composition of the population used to study it (Xie et al., 2012). Understanding variation in responses to performance pay can thus aid in tailoring compensation systems to particular organizations composed of particular sorts of workers. The current focus on sorting effects is a further limitation. While it is important for organizations to know the types of employees they are likely to attract given their compensation systems, it is equally important for them to know how their current workforce is likely to react to changes in how they are paid.

According for Different Forms of Performance Pay

A further complication arises due to the fact that although differential responses to incentives writ large may be the norm, what these responses look like may also vary according to the form of performance pay offered. That is, just as not all employees are likely to respond in
the same way to the same incentive, the same employees may respond in yet differ ways when they are offered different types of incentives. Piece rates, merit pay, and gainsharing, for instance, may have different relationships to worker ability and therefore may each yield a different distribution of employee responses.

A key point of differentiation among performance pay systems is whether they reward outcomes tied to individual or group effort (Gerhart et al., 2009), and here too ability should shape the employee response because collective rewards alter the connection between effort and reward and because they introduce new interpersonal dynamics (Prendergast, 1999). An obvious change under group incentives is that an employee’s impact on collective output and thus the reward she receives decreases as group size increases. This creates opportunities for freeriding, where an employee receives the benefits of her coworkers’ increased efforts even if she does not increase her own. There is also a greater role for social influences, like peer pressure and knowledge sharing, under group performance pay. In terms of incentive effects under group performance pay, then, they should be more powerful among the less abled because the less able are more susceptible to the social influence of the more able and because the more able may feel that they will not benefit in proportion to the added effort they expend. In terms of sorting effects, a preference for group performance pay should be greater among the less able because they can benefit from the efforts of their more able coworkers.

The research on differential response to group performance pay is limited as well. For sorting effects, there is mixed evidence on the relationship between ability and performance pay preferences. Trank et al. (2002) find that high performing students prefer individual pay. In a survey of U.S. manufacturing firms, Park et al. (1994) find a positive relationship between the turnover of top performers and group incentives and a positive relationship between the turnover
of poor performers and individual incentives. These findings conflict with those of Hamilton et al. (2003), who, looking at actual sorting in a firm that switched from individual to group piece rates, find that more productive employees were more likely to choose to work in teams and that turnover under group performance pay was no different among the more and less able.

The little evidence there is on incentive effects has more consistent support. In a lab study, Beersma et al. (2003) find that group performance pay has a bigger impact on the less able than the more able. Under group incentives, low performers in their experiment increased their accuracy when completing their task more than top performers did. Looking at firm data, both Weiss (1987) and Hansen (1997) find that group performance pay improves the performance of the less productive and reduces that of the more productive. The incentive effect under group performance pay thus appears stronger among the less able.

Taken together, the observations in this section highlight the value of examining 1) the variation in individuals of different ability levels’ responses to the same incentive, and 2) how this variation relates to the type of incentive they are offered. These observations further reveal that adequately investigating both points requires researchers to distinguish between incentive and sorting effects. I now turn to a field experiment that improves on past work.

**SETTING**

The shipping division at PickInc\(^2\), a large, U.S. retailer, provides the setting for this study. PickInc sells products to businesses and consumers at retail stores, online, and through contracts.

---

\(^2\) The name has been changed to preserve confidentiality.
Online orders are processed at warehouses called fulfillment centers (FCs), five of which served as the main research sites. FCs comprise about 10 departments. Each department is responsible for one task, and all non-supervisory members of the department perform the same task.

The flow of work within the FCs is akin to an assembly line. Departments do not directly coordinate with each other, but the pace of work in one department can affect the pace of work downstream. Workers are typically attached to one department, although some are cross-trained, which means they can move to other departments to help reduce bottlenecks when their managers tell them to do so. A workday is intended to last for eight hours, but workers must stay on if there is more work to be done. Workflow varies with order volume, which tends to peak early in the week before slowing on Friday.

In 2017, PickInc managers decided to pilot a performance pay plan in the picking department, which employs the largest share of workers within the FCs. My role was to evaluate the effectiveness of this pilot. My work included suggesting an appropriate research design and data collection procedures, and I conducted the data analysis. In addition to looking at administrative records already kept internally at PickInc, this second task involved conducting interviews with workers and managers, conducting surveys, and participating in regular project planning calls. (These elements are described below in more detail.)

The central task in the picking department is to “pick,” or retrieve objects and place them into boxes. There are multiple ways to pick, but management opted to implement the pilot in FCs that all utilize what is called “pick-to-cart” technology. Under this method, workers grasp items from shelves arrayed throughout the FC and place them into boxes that sit on large, wheeled carts. Every item put in a box is called a line. Some boxes have many lines and some boxes have few. Likewise, some items are small, such as a USB thumb drive, and others are big, such as a
case of soda. Each worker, called a "picker," is responsible for her own cart, which holds 40 to
50 boxes. When the picker fills all her boxes, she obtains a new cart. She is guided on an optimal
route through the warehouse by an electronic headset. Crucial to this study’s ability to identify an
effect of performance pay, the boxes on her cart are delegated by a computer system, leading to
random assignment of orders over time. A typical day involves pulling around seven carts.
Pickers are expected to pick at a rate of 120 to 130 lines per hour (LPH). Picking is the entry
level job at the FC and takes about three months to master.

Pick-to-cart picking creates a low level of interdependence among workers. PickInc uses
what is a called a “batch” picking strategy, in which a single picker is responsible for filling an
entire order from start to end. This contrasts with more interdependent alternatives, such as
“zone” picking, where orders are passed from one picker to the next (Tompkins, White, Bozer,
& Tanchoco, 2003). In support of the presumption of low interdependence, a pre-incentive survey
of pickers showed that 89% of the 148 respondents agreed or strongly agreed with the statement,
“I can do my job with minimal help from others.” That said, although a cart requires only one
person to be filled, interference and assistance is possible. Slower carts, for example, get in the
way of faster ones, which can slow the whole picking process, leading to “traffic jams.”
Moreover, workers may help push their coworkers’ carts when they get very heavy, and they
may share their knowledge about location of items and picking techniques.

PickInc devised two different types of performance pay for it pickers. Under an
individual plan, a pickers’ pay depends on her own ability to meet a set of tiered LPH standards
set by PickInc’s engineering department. Under a group plan, a pickers’ pay depends on her
department’s ability to meet an analogous set of standards. The performance tiers are
incremented at 5 percent increases in LPH, starting at the baseline of 130 LPH, and are based on
building-specific standards that are designed to allow for comparison across sites. As she reaches each tier, the worker receives an additional $0.25 per hour, for up to $1.50 per hour. Figure 1 portrays this tier structure. No matter her performance (i.e., she picks below 130 LPH), a worker earns a minimum wage corresponding to the flat wage she received prior to the pilot.³ As an example, consider a picker who worked 35 hours per week each month and picked at an average rate of 135 LPH before the pilot. She took home $2,100 each month. Under the individual plan, she will take home $2,170, or an additional $70, if she starts working at a rate of 145 LPH.

The incentives are paid according to performance over a company-defined period, which is roughly one month long. To qualify, workers must meet pre-established quality standards and attendance requirements. Workers in any department are eligible, but the payout they receive applies only to hours spent picking, and they must work for at least one hour total in the picking department per period to be eligible.

³ PickInc has not implemented a comparable incentive plan before. Past efforts at rewarding workers were site-specific, low-value (i.e., $25 or less gift cards), and short-lived.
In designing these plans, several features of picking and FC operations figured in PickInc managers’ decision. Pickers are primarily evaluated by their picking productivity, and this measure is recorded electronically. Individual performance from the previous day is also publicly displayed. (This is true both at the FCs receiving the incentives and those not.) Managers thus decided to use this objective, public measure of performance in both incentive plans. Recall that the variation in incentive types stems from the fact that the two incentives are tied to different outcome levels. On the one hand, managers wanted to exploit their ability to track individual performance by offering an individual incentive. On the other hand, they wanted to instill a cooperative attitude among their workforce and felt that a group incentive could achieve this result. Piloting two incentive types was seen as a way to test the relative strengths of each in the FCs. PickInc managers said they would deem the pilot successful if they witnessed a 10 percent productivity gain in the picking department under either plan.

HYPOTHESES

Overall, I expect both incentives to boost picker productivity. The individual plan should yield a greater average increase in productivity than the group plan because picking can be performed with minimal help from others and because individual performance is readily measured. Picking is low in task interdependence, which implies that increased cooperation will have a weak effect on collective output and thereby on the payout pickers receive (Wageman & Baker, 1997). Countervailing reasons in favor of the group incentive come in two forms. One, the group incentive may induce cooperative behaviors. Such behaviors are plausible in this setting because performance is observable and because the fairly small group size weighs against
freeriding. A single picker exerts identifiable influence on collective output and so she may be more readily held accountable if she shirks (Wageman & Baker, 1997). Second, the individual plan may backfire, leading to dysfunctional, uncooperative behavior. A main reason for this is that it will create greater pay dispersion among pickers, rendering the plan more susceptible to negative social comparisons (Larkin, Pierce, & Gino, 2012).

Turning to differential responses, this paper’s focus, I expect the positive effects under the individual incentive to be stronger among those who are already performing at high levels (H1). This is because it will take them relatively less effort to increase their productivity (Lazear, 2000). In other words, the relationship between the incentive and productivity will increase with ability. Stark and Hyll (2011) provide a counter to this hypothesis: because high performing workers stand to gain more than low performers, feelings of relative deprivation could actually motivate the low performers to expend more effort.

On the other hand, I expect the positive effects under the group incentive to be stronger among low performers (H2). This is because they are likely to feel more social pressure to improve their performance so that their department can secure a payout (Kandel & Lazear, 1992). In other words, the relationship between the incentive and productivity will decrease with ability. The opposite outcome could occur if less able workers are more likely to shirk, leaving the high performers to carry most of the load (Holmstrom, 1982). The fact that performance is publicly known weighs against this outcome.

RESEARCH DESIGN AND DATA
I explore these hypotheses via a field experiment that generated a variety of data. As mentioned earlier, PickInc tracks production data, including output and quality, for each worker each day. The company also keeps detailed personnel records containing information like pay, attendance, and turnover, which I was able to access.

The experiment proceeded as follows. PickInc managers determined which FCs would participate in the pilot. They prioritized selecting comparable facilities. Because there are multiple picking methods, their choice to focus on pick-to-cart technology significantly narrowed the pool. One FC was excluded because the laws in the state where it was located made it more difficult to implement an incentive plan. Two more were deemed too distant from the researcher and the lead managers for the multiple visits required by the study. This left six buildings to participate, one of which was excluded from this study so that it could take part in a follow up incentive trial held at a later date.

It was not possible to vary incentive type within the FC. The managers selected two out of the five FCs to receive “treatment,” that is, serve as candidates for the incentive, and I then randomly assigned the incentive type to these buildings such that all workers in them were given the same incentive. The remaining three FCs, which were held by the managers to be operated competently relative to other FCs in the network and comparably to the treated FCs, served as “controls.” Figure 2 shows the treatments each FC received.

Figure 2: Incentive Treatments

<table>
<thead>
<tr>
<th></th>
<th>Individual</th>
<th>Group</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligibility</td>
<td>Individual minimum LPH</td>
<td>Group minimum LPH</td>
<td>130 LPH 143 LPH</td>
</tr>
<tr>
<td># FCs</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
Three points about treatment assignment are worth noting. First, the LPH standards are site-specific. This reflects the results of engineering time studies carried out in each FC. While the picking methods are the same in participating FCs, slight variations in building dimensions and layouts yield different performance standards. Second, the nature of PickInc’s industry means that seasonal fluctuations in order volume are felt equally across its FCs, even though they are in distant locations. Third, orders are directed to FCs based on customer location. Although FCs do share inventory when certain items are out of stock somewhere in the network, PickInc’s order allocation process means that customer-driven changes in demand in a given region are unlikely to be re-routed to more distant FCs. As a result, on-the-ground changes in a given FC, including those caused by this experiment, are unlikely to affect other FCs in the network. (Unexpected major disruptions can lead orders to be re-routed, but this did not occur during the course of the pilot.) These points support comparability. Moreover, in what follows I employ an estimation strategy that should help overcome initial differences in FC conditions.

A launch date, the same for both incentive types, was chosen to coincide with PickInc’s internal calendar, not with any particular segment of the business cycle. Three months prior to this date, I began compiling daily production data. Treated workers were told of the pilot plan by their direct supervisors one week before it launched. The pilot then ran for three months, giving the workers three opportunities to qualify for extra pay. The workers were told of this duration but were also told there would be opportunity to continue with the plan after assessing the results and incorporating their feedback.

Before the incentive was announced, I administered a survey to the workers in the picking departments at each FC. I also interviewed pickers, a handful of workers in other departments, and several supervisors in person at the FCs. At the end of the pilot, I administered
a second survey to the treated FCs. (These data were valuable in developing and assessing the performance pay programs, but they will be discussed in a future paper and so are not relied on at length in the current study.)

DESCRIPTIVE STATISTICS

Analysis of data prior to the launch of the pilot does not reveal any systematic differences between the participating FCs, although it does suggest some dissimilarities. Table 1 shows that the FCs are comparable in size and are located in different regions of the U.S.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Region</th>
<th>Number of Pickers*</th>
<th>Number of Workers</th>
<th>Average Annualized Turnover Rate^t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Southwest</td>
<td>29</td>
<td>86</td>
<td>47% (58%)</td>
</tr>
<tr>
<td>Group</td>
<td>South</td>
<td>51</td>
<td>151</td>
<td>99% (79%)</td>
</tr>
<tr>
<td>Control 1</td>
<td>Mid-Atlantic</td>
<td>27</td>
<td>94</td>
<td>104% (119%)</td>
</tr>
<tr>
<td>Control 2</td>
<td>Northeast</td>
<td>27</td>
<td>77</td>
<td>42% (60%)</td>
</tr>
<tr>
<td>Control 3</td>
<td>Midwest</td>
<td>53</td>
<td>200</td>
<td>83% (60%)</td>
</tr>
</tbody>
</table>

Note: Both the number of pickers and workers were counted at the start of the pre-treatment period. *To be classified as a picker, one had to be a designated member of the picking department. ^The annualized turnover rate is calculated at the building level each week and then averaged for the 6-months prior to the start of the pre-treatment period. Standard deviations in parentheses.

Operations are similar across the FCs in this study. In the picking departments, there are two shifts. One starts in the late morning, the other in midafternoon. Pickers in the second shift are responsible for filling all outstanding orders and cannot leave until they do so. Overtime is a common occurrence. Pickers, for the most part, say they interact with those on both shifts and do not differentiate between full-time and part-time co-workers in terms of personal relationships. Temps, which were not allowed to participate in the pilot and so are kept out of the data, are sometimes hired to fill gaps. Permanent employees say they tend not to forge bonds with temps.
Table 2 captures characteristics of the workforce prior to the launch of the pilot. Looking at this table, it is important to note once again that any worker who picks for a total of more than one hour each period in the treated FCs is eligible for an incentive. Because workers from other departments pick when their supervisors request that they do so, I observe workers from departments other than picking with some frequency. That said, Table 2 shows that around 80 percent of the workers who picked in the pre-treatment period were official members of the picking department. As for experience, the average worker has been with the firm for a little more than a year. Turnover fluctuates from month-to-month, but managers have indicated that reducing turnover is a priority at every FC.

<table>
<thead>
<tr>
<th></th>
<th>Individual</th>
<th>Group</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Share in picking department (%)</td>
<td>83.58</td>
<td>86.95</td>
<td>79.64</td>
</tr>
<tr>
<td></td>
<td>(37.09)</td>
<td>(33.68)</td>
<td>(40.27)</td>
</tr>
<tr>
<td>Share full-time (%)</td>
<td>65.97</td>
<td>92.15</td>
<td>78.67</td>
</tr>
<tr>
<td></td>
<td>(47.40)</td>
<td>(26.91)</td>
<td>(40.97)</td>
</tr>
<tr>
<td>Tenure (years)</td>
<td>1.41</td>
<td>0.97</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td>(3.70)</td>
<td>(2.42)</td>
<td>(3.76)</td>
</tr>
<tr>
<td>Hourly wage ($)</td>
<td>14.42</td>
<td>15.46</td>
<td>15.79</td>
</tr>
<tr>
<td></td>
<td>(0.84)</td>
<td>(0.72)</td>
<td>(1.37)</td>
</tr>
<tr>
<td>Number of observed workers</td>
<td>32</td>
<td>32</td>
<td>106</td>
</tr>
</tbody>
</table>

Note: These data represent averages over the pre-treatment period. Some records were not available for all workers present in the dataset that is used to calculate the treatment effect in the main analysis. This was the case for only a small fraction (< 0.05) of observations in participating buildings, except for one control, which was missing a substantial share. In each building, missing picking department status was imputed based on the typical total hours spent picking in each department in that building; members of some departments are more likely to log picking hours than others. Full-time status was imputed based on mean hours spent picking. Excluding the imputed data has a negligible impact on these numbers.

The pay structure is fairly flat within the FCs. Pickers in this study all earn between $14 and $15 per hour. Within FCs, there is little variation in the picking wage. Any gaps in pay in

---

4 I refer to any worker undertaking picking as a "picker," even if she is formally assigned to a different department.
5 Picking is the entry level job at the FCs. Because the majority of workers in this study are pickers, their tenure with PickInc should largely reflect their time in picking.
picking departments in the same building are due to seniority: workers receive a small increase in wages each year for a fixed number of years before maxing out. Workers in other departments in the same building may earn more because they have received training to operate certain pieces of machinery. Across FCs, pay varies in line with local costs of living. PickInc managers judge these wages to be comparable to competitors’, and many workers say it is satisfactory.

While data on worker demographics are not available, visits to the FCs produced several observations that provide insight into this aspect of the setting. Men and women are about evenly represented in the picking department, with the second group being perhaps a bit larger. There is diversity in race and ethnicity, but the makeup of the workforce appears to be reflective of the region in which the FC is located. All age groups are represented, though there are fewer older pickers because of the physical nature of the job. Workers in the FCs describe themselves as not being highly educated. Some of the younger ones, however, are either planning on returning to school or taking classes part-time.

The initial surveys show that pickers are generally satisfied with their situation. In particular, they say their relations with each other are positive. For the most part, they are friendly to their coworkers, with many offering the word “family” to describe them.

METHOD

Outcomes

PickInc’s primary goal was to boost productivity, and productivity is this study’s focal outcome. Productivity is measured in lines per hour (LPH). PickInc tracks individual LPH on a
daily basis using a computer system. The key inputs in this measurement are the hours a picker is logged into the picking department and the number of lines she picks in that time.

To make up for potential measurement errors, I restrict the sample according to a rule of thumb suggested by PickInc managers. Each week supervisors are expected to correct any errors in record-keeping that could generate incorrect LPH readings, though this does not always happen. Errors typically occur when pickers move to other departments without properly badging in and out. Any extremely high LPH readings (> 1,000) can, for the most part, be deemed mistakes. Remaining readings above 300 LPH are plausible but probably unreliable. Only four readings out of 16,295 initial observations exceeded 300 LPH. Excluding them reduces the uncertainty of my estimates. It does not change my general conclusions.

In addition, the daily hours a worker spends picking is sometimes improperly recorded. I cut observations where this measure was greater than 14.5 hours. Such readings likely stem from mistakes clocking in and out, as I triangulated them with payroll data used to actually generate paychecks and found that no workers actually exceeded this hours threshold. This resulted in 16 observations being discarded, leaving a total sample of 16,275 employee-day observations.

I then restrict the sample to workers present at least once in both the pre-incentive and incentive periods. This is done so that each worker’s relative ability, described more below, can be tracked in both periods. If a picker is observed only once in a period, it is likely that he or she is a member of a different department who was asked to help pick during a one-off event. As a result of this winnowing, the sample drops to 14,550 observations, composed of 170 pickers. This sample is the basis of the tables and analyses used throughout the paper.

Finally, in cases where I want to rule out sorting effects, I drop observations of any picker who quit the firm or was hired in the post-incentive period. New hires are already excluded from
this sample by the above steps. Exits occurred 7 times in the FC receiving the individual incentive, 5 times in the FC receiving the group incentive, and 9, 7, and 4 times in the control FCs. This yielded a sample of 12,985. I explore how entry and exit affects the results below.

**Moderators**

Based on my above hypotheses, ability should moderate the incentives’ effects on productivity. Underlying ability is hard to observe, so I follow Lazear (2000) in focusing on actual output. Ability is hence measured as a picker’s average productivity in the three months prior to the incentive. This measure assumes that observed output is a proxy for ability. It has the benefit that it is not self-reported. Figure 3 displays the spread of daily productivity readings by FC. It shows that ability roughly follows a normal distribution and is comparable across FCs.

![Figure 3: Pre-Incentive Distribution of Daily LPH, By Treatment Type](image)

*Note: Dotted lines represent mean productivity readings by FC. The average for the controls is 117.55; the individuals is 121.32; and the group is 118.99. These histograms have been normalized so that each group’s distribution can be more readily compared given that the samples are of different sizes.*
However, an essential point in the existing theorizing and empirical research on the relationship between incentives and ability is that ability is a relative concept. Because incentives are often offered to whole teams where interpersonal relations and comparisons are at play, it makes more sense to analyze relative rather than absolute ability. To measure relative ability, I sort pickers into deciles within their FC based on their average productivity levels before the incentive period. This variable serves as the study’s main moderator. Table 3 displays the rankings. Importantly, it shows that in both FCs receiving an incentive, only pickers in the top decile were performing at levels that made them eligible for a payout in the pre-incentive period.

<table>
<thead>
<tr>
<th>Decile</th>
<th>Individual</th>
<th>Team</th>
<th>Control 1</th>
<th>Control 2</th>
<th>Control 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64.74</td>
<td>95.22</td>
<td>93.08</td>
<td>82.05</td>
<td>85.26</td>
</tr>
<tr>
<td>2</td>
<td>108.57</td>
<td>99.46</td>
<td>112.25</td>
<td>96.90</td>
<td>103.57</td>
</tr>
<tr>
<td>3</td>
<td>111.54</td>
<td>102.89</td>
<td>115.44</td>
<td>111.63</td>
<td>105.87</td>
</tr>
<tr>
<td>4</td>
<td>115.17</td>
<td>108.03</td>
<td>118.41</td>
<td>121.33</td>
<td>110.05</td>
</tr>
<tr>
<td>5</td>
<td>118.61</td>
<td>115.25</td>
<td>120.38</td>
<td>124.03</td>
<td>114.45</td>
</tr>
<tr>
<td>6</td>
<td>122.89</td>
<td>121.13</td>
<td>121.42</td>
<td>126.29</td>
<td>117.87</td>
</tr>
<tr>
<td>7</td>
<td>125.22</td>
<td>125.87</td>
<td>123.03</td>
<td>130.84</td>
<td>119.30</td>
</tr>
<tr>
<td>8</td>
<td>127.34</td>
<td>129.78</td>
<td>126.07</td>
<td>136.84</td>
<td>122.72</td>
</tr>
<tr>
<td>9</td>
<td>134.66</td>
<td>132.67</td>
<td>135.07</td>
<td>144.03</td>
<td>128.18</td>
</tr>
<tr>
<td>10</td>
<td>146.61</td>
<td>147.82</td>
<td>167.60</td>
<td>157.72</td>
<td>136.35</td>
</tr>
</tbody>
</table>

**Table 3: Pre-Treatment LPH Decile Rankings, by FC**

**Estimation Strategy**

To test my predictions about worker productivity, I employ a difference-in-differences (DD) strategy that uses pre- and post-treatment information. The DD takes the following form:

\[ Y_{ijt} = \alpha_1 Individual_{jt} + \alpha_2 Group_{jt} + \gamma_i + \tau_t + \varepsilon_{ijt} \]
where \( i \) indexes workers in FC \( j \) and \( t \) indexes day, and \( \gamma_i \) and \( \tau_t \) are, respectively, fixed effects for worker and day, where day corresponds to the calendar date. The dependent variable (LPH) is \( Y_{ijt} \), a variable bounded at zero. The binary indicator variables \( \text{Individual}_{jt} \) and \( \text{Group}_{jt} \) take the value of 1 in the FCs receiving these incentive types on days after the launch of the pilot. The coefficients of interest, or treatment effects, are \( \alpha_1 \) and \( \alpha_2 \), which correspond to the different incentive types. I expect that \( \alpha_1 \) should be greater than \( \alpha_2 \) and that both should be positive.

DD estimation requires that the FCs chosen for treatment be compared to similar FCs that do not receive treatment. This strategy provides an unbiased estimate of the causal effect of interest so long as it meets the “parallel trends” assumption that the treated FCs would have experienced the same change in outcomes over time as the untreated FCs had they not received treatment. In this setting, the parallel trends assumption is plausible because the tasks are the same across the FCs, the composition of the workforce and of management are comparable, and the FCs face similar fluctuations in demand. Although I evaluate the “parallel trends” assumption below, I can employ a weaker form of it called the “parallel growths” assumption (Mora & Reggio, 2013), as in the following equation.

\[
Y_{ijt} = \alpha_1 \text{Individual}_{jt} + \alpha_2 \text{Group}_{jt} + \eta_jT + \gamma_i + \tau_t + \epsilon_{ijt}
\]

Here, \( \eta_jT \) are FC-specific linear time trends. Including these trends means that I no longer have to assume that the FCs would have followed similar paths in the absence of treatment.

These regression equations yield estimates of the average effects of the incentive, but we are really interested in the variability of these effects. We can assess how the pickers’ responses vary according to ability by performing what is known as subgroup analysis. When we do so, we are no longer estimating the average treatment affects (ATEs) of the incentives but the
conditional average treatment effects (CATEs) because the effects are now conditioned on worker ability (Gerber & Green, 2012).

There are two ways to model CATEs in a DD framework. The first assumes that ability has a linear relationship with treatment and takes the following form:

\[
Y_{ijt} = \alpha_1 Individual_{jt} + \alpha_2 Group_{jt} + \beta_1 (Individual_{jt} \times Ability_i) + \beta_2 (Group_{jt} \times Ability_i) + \eta_j T + \gamma_i + \tau_t + \epsilon_{ijt}
\]

where \(Ability_i\) is a continuous variable corresponding to each picker’s average pre-treatment productivity, measured in LPH. The CATE is estimated for each incentive type by summing the corresponding coefficients \(\alpha\) and \(\beta\). Hypothesis \(H1\) predicts that this sum should be positive for both incentive types, and hypothesis \(H2\) predicts that the sign on \(\beta_1\) should be positive while the sign on \(\beta_2\) should be negative.

The above model assumes that the effect of incentives on productivity, conditional on ability, changes at a constant rate given by the \(\beta\)’s. This may not be a good assumption, for two reasons. First, it may be that the size of the incentive effect changes with ability at a non-constant rate. For example, in the event that the least able pickers don’t respond to individual incentives at all, medium ability pickers respond positively and strongly, and high ability pickers respond positively but weakly, this relationship would be missed. Second, if we lack enough observations at certain ranges of the ability distribution, we may arrive at biased estimates of the incentives’ effects (Hainmueller, Mummolo, & Xu, 2018). We can get around this potential misstep by assuming that the effects are nonlinear, as in the following form:

\[
Y_{ijt} = \alpha_1 Individual_{jt} + \alpha_2 Group_{jt} + \beta_{11} Individual_{jt} Rank_1 + \cdots + \beta_{1n} Individual_{jt} Rank_n + \beta_{21} Group_{jt} Rank_1 + \cdots + \beta_{2n} Group_{jt} Rank_n + \eta_j T + \gamma_i + \tau_t + \epsilon_{ijt}
\]
Here I sort pickers into 10 bins according to their decile rankings of average pre-incentive productivity, and \( \text{Rank}_n \) is now a dummy variable that takes 1 when an individual’s pre-treatment decile ranking is equal to \( n \). Each incentives’ effect on those in the first decile is estimated as the \( \alpha \)’s. For those in higher rankings, the effect is estimated by summing the corresponding \( \alpha \) with \( \beta_n \) where \( n \) is equal to their pre-incentive decile. Hence, this model compares pickers in different treatments yet similar places within the ability hierarchy. For the individual incentive, I hypothesize that coefficient \( \beta_{1n} \) should be increasing with \( n \). For the group incentive, I hypothesize that coefficient \( \beta_{2n} \) should be decreasing with \( n \). I do not make predictions about the relationship between specific deciles.

Note that the treatment assignment process and the DD approach imply that in all specifications I should not see substantive changes to my estimates when I include control variables. Nevertheless, a few points are worth making in light of concerns with the site-specific features of the FCs and their workers. First, any of the unchanging workers characteristics that may influence the incentives’ effectiveness, including race or gender, will be swept away by the worker fixed effects. The same goes for any features unique to the FC that affects all workers. To alleviate any concerns that differences in business activity could be affecting productivity, in some specifications I include a variable tracking daily volume (the total number of boxes ordered) in the picking department.

The coefficients in all models are estimated using the \texttt{felm} command in R’s \texttt{lfe} package. Standard errors are estimated using Stata’s \texttt{reghdfe} command, which includes a degrees-of-freedom adjustment and a small-sample correction.
RESULTS

Overall Effects

The initial results indicate that only the individual incentive boosted picker productivity. Table 4 shows that LPH increased at the FC receiving the individual incentive in the three months that the pilot ran, while it fell at the FC receiving the group treatment. There was little change at the controls. According to managers, productivity usually declines during the time of year in which the pilot was run because PickInc’s order volume is lower. The results translate into a 4.2% increase in average productivity under the individual incentive and a 3.5% decline under the group incentive.

Table 4: Unconditional Mean LPH by Treatment Type

<table>
<thead>
<tr>
<th></th>
<th>Pre-Treatment</th>
<th>Post-Treatment</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>121.32</td>
<td>126.44</td>
<td>5.12***</td>
</tr>
<tr>
<td></td>
<td>(27.92)</td>
<td>(30.55)</td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>118.99</td>
<td>114.79</td>
<td>-4.21***</td>
</tr>
<tr>
<td></td>
<td>(21.82)</td>
<td>(19.13)</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>117.55</td>
<td>118.30</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>(20.06)</td>
<td>(19.82)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard deviation in parentheses. *p<0.05; **p<0.01; ***p<0.001

Figures 4 and 5 depict these results graphically. The first figure shows the distribution of worker productivity before and after the pilot for each treatment type. The second shows the daily difference in mean worker LPH between the treated and control FCs. Pickers in the FC receiving both treatments were slightly more productive than those at the control FCs before the pilot began. During the pilot, those in the individual treatment became much more productive than the controls. The pickers receiving the group incentive declined in standing. Importantly, the productivity differences between the treated and control FCs were fairly steady before the pilot, as illustrated by Figure 5. This is essential for the parallel trends assumption to be met.
Figure 4: Change in Distribution of Daily LPH, By Treatment Type

Change in Productivity Distribution for Controls

Change in Productivity Distribution Under Individual Incentive

Change in Productivity Distribution Under Group Incentive
Figure 5: Difference in Mean Daily LPH between Treated and Controls, By Treatment Type

Note: The solid lines show the mean differences in productivity for each incentive type compared with the control, before and after treatment. The vertical line marks when the pilot launched, and the yellow shading marks when the pilot was announced.

Moving to the DD results, statistically significant and positive treatment effects are apparent under the individual incentive. In Table 5, Models 5 and 6 are preferred. The first of these includes worker and day fixed effects along with order volume as a control. The second adds the FC-specific time trend. The individual treatment effects remain significant in each specification. On the other hand, the group incentive produced lackluster results. In Table 5, the effect is negative in every specification. Although the effect varies in significance, it seems reasonable to say that this treatment did not boost picking performance.
Table 5: DD Regression: Effect of Incentives on Productivity

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>4.374*</td>
<td>4.303*</td>
<td>5.844***</td>
<td>5.849***</td>
<td>5.931***</td>
<td>4.629*</td>
</tr>
<tr>
<td></td>
<td>(2.119)</td>
<td>(2.156)</td>
<td>(1.577)</td>
<td>(1.596)</td>
<td>(1.586)</td>
<td>(2.178)</td>
</tr>
<tr>
<td>Group</td>
<td>-4.953</td>
<td>-4.914</td>
<td>-7.319**</td>
<td>-7.341**</td>
<td>-7.263**</td>
<td>-0.495</td>
</tr>
<tr>
<td></td>
<td>(2.594)</td>
<td>(2.610)</td>
<td>(2.328)</td>
<td>(2.361)</td>
<td>(2.359)</td>
<td>(1.879)</td>
</tr>
<tr>
<td>Department Volume</td>
<td>-0.0003</td>
<td>-0.001*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0005)</td>
<td>(0.0004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual fixed effects? No No Yes Yes Yes Yes
Time fixed effects? No Yes No Yes Yes Yes
Building-specific time trends? No No No No No Yes
Observations 14,550 14,550 14,550 14,550 14,550 14,550

Note: *p<0.05; **p<0.01; ***p<0.001
DV is daily lines per hour (LPH). Standard errors clustered by individual in parentheses.

The individual treatment effect is substantive as well. Taking the coefficient in Model 6 as an example, the individual incentive boosted worker productivity by about 5 lines per hour. A box holds, on average, about 4 lines. Hence, a worker who picked for 7 hours a day would fill 8 more boxes under the individual incentive than she did prior to the pilot.

To parse out sorting effects, I rerun this regression on two different samples. Table 5 relies on the sample with only those who were present in both pre- and post-incentive periods. Table A1, in Appendix A, displays the results for the entire sample of workers who picked for at least an hour each period. Here, the coefficient on the individual effect in the full specification (Model 6) is about 40% higher than that in Table 5. The group coefficient, though positive, is insignificant. Because the coefficient increases under both incentives in this sample, which includes new hires, this provides suggestive evidence that the entrants were more able. Table A2 is more restrictive. It uses a sample that excludes both those who exited the firm or were hired during the incentive period. The coefficient on the individual effect in the full specification (Model 6) is now about 28% higher than that in Table 5, and the coefficient on the group effect
has grown more negative, though it is still insignificant. This suggests that the individual incentive may have caused poor performers to leave, whereas the group incentive may have caused better performers to leave.

While there were measurable changes to productivity under the incentives, the actual payouts were limited. None were received in the group treatment because productivity fell. Only six pickers received a boost in monthly pay under the individual treatment, although 23 times a picker actually qualified for the bonus but was later disqualified for attendance violations. Of the qualifiers, only one was performing well enough before the incentive period that she would have earned a payout even if she did not increase her performance. The average size of the payout was $84.00, with a peak of $178.35.

**Heterogeneous Treatment Effects**

One reason for the finding that the individual incentive had demonstrable effects on behavior even if few payouts were received could be that workers of differing abilities responded to the incentives in different ways. By the same token, some pickers could have increased their performance under the group incentive even if their efforts were not strong enough to shift their department above the minimum required for a payout. Consequently, it is worth investing the heterogeneous responses to the incentives.

Table 6 treats ability as a continuous moderator of the incentive response. It reveals that the effects of both incentive types decrease as one’s ability increases. Figures 6a and 6b show these results graphically. Both incentives initially have a positive effect on productivity, but this effect decreases in strength as one moves up the ability rankings.
Table 6: Continuous Heterogeneous Treatment Effects

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>11.789*** (2.781)</td>
<td>10.428*** (2.635)</td>
</tr>
<tr>
<td>Group</td>
<td>11.743** (3.585)</td>
<td>17.816*** (2.878)</td>
</tr>
<tr>
<td>Individual x Rank</td>
<td>-1.035* (0.464)</td>
<td>-1.056* (0.464)</td>
</tr>
<tr>
<td>Group x Rank</td>
<td>-3.113*** (0.491)</td>
<td>-3.071*** (0.495)</td>
</tr>
<tr>
<td>Volume Control?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Individual fixed effects?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Time fixed effects?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Building-specific time trends?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>14,550</td>
<td>14,550</td>
</tr>
</tbody>
</table>

Note: *p<0.05; **p<0.01; ***p<0.001
DV is daily lines per hour (LPH). Standard errors clustered by individual in parentheses.

There are two takeaways from these figures. First, the individual incentive boosted productivity at nearly all ability levels, though its effect diminished. This supports my initial expectations on overall incentive effects, but it runs counter to hypothesis H1, which anticipated that the individual incentive would be more attractive to top performers. Second, while the group incentive failed to boost overall productivity, it did have a positive effect on pickers in between the first and fourth to sixth deciles. This suggests that around half of pickers did respond favorably to the group incentive. It also supports hypothesis H2, which claimed that high performers would not be positively moved by the group incentive.
Figure 6: Linear Marginal Effect of Incentives on Productivity Moderated by Ability

Figure 6a: Without building specific time trends.

Figure 6b: With building specific time trends.
Finally, I examine treatment effects without assuming they are linear. Recall that the weakest pickers, those in the first decile, serve as the reference category. To estimate the effect on a given decile, we add the coefficient for that decile in Table 7 to the estimate for the first decile. F-tests confirm that the effects vary by rankings under both treatments ($p < 0.01$). Table 7 shows consistent, significant decreases in the group incentive’s effect as ability increases.

<table>
<thead>
<tr>
<th>Table 7: Nonlinear Heterogeneous Treatment Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
</tr>
<tr>
<td><strong>Individual (Rank = 1st Decile)</strong></td>
</tr>
<tr>
<td>Individual x 2nd Decile</td>
</tr>
<tr>
<td>Individual x 3rd Decile</td>
</tr>
<tr>
<td>Individual x 4th Decile</td>
</tr>
<tr>
<td>Individual x 5th Decile</td>
</tr>
<tr>
<td>Individual x 7th Decile</td>
</tr>
<tr>
<td>Individual x 8th Decile</td>
</tr>
<tr>
<td>Individual x 9th Decile</td>
</tr>
<tr>
<td>Individual x 10th Decile</td>
</tr>
<tr>
<td><strong>Group (Rank = 1st Decile)</strong></td>
</tr>
<tr>
<td>Group x 2nd Decile</td>
</tr>
<tr>
<td>Group x 3rd Decile</td>
</tr>
<tr>
<td>Group x 5th Decile</td>
</tr>
<tr>
<td>Group x 6th Decile</td>
</tr>
<tr>
<td>Group x 7th Decile</td>
</tr>
<tr>
<td>Group x 8th Decile</td>
</tr>
<tr>
<td>Group x 9th Decile</td>
</tr>
<tr>
<td>Group x 10th Decile</td>
</tr>
</tbody>
</table>

| Volume Control? | Yes | Yes |
| Individual fixed effects? | Yes | Yes |
| Time fixed effects? | Yes | Yes |
| Building-specific time trends? | No | Yes |
| Observations | 14,550 | 14,550 |

Note: *$p<0.05$; **$p<0.01$; ***$p<0.001$

DV is daily lines per hour (LPH). Standard errors clustered by individual in parentheses.

33
The picture is more complicated for the individual incentive. This relationship is easier to see when visualized, as in Figures 7a and 7b. They display a J-shaped curve for the individual incentive and a downward sloping curve for the group incentive. Hence, the individual incentive has the most positive effect on those of lowest ability, no effect on the median picker, and a smaller positive effect on the most able pickers. This adds a check on our conclusions surrounding hypotheses H1 and H2 that would not be uncovered without modeling heterogeneous effects. That is, individual incentives appear to be attractive to both top and bottom performers but not middle ones.

**Figure 7:** Nonlinear Marginal Effect of Incentives on Productivity Moderated by Ability

---

**Figure 7a:** Without building specific time trends.
To see if these results are due to sorting, I perform several additional tests. First, I use the same sample as above but rank pickers according to their standing in the full sample. (The results are displayed in Appendix A in Figure A.1) This is done in case the restricted sample makes pickers appear more or less able than they really are. Second, I use a sample that contains no new hires or exits during the incentive period. I again rank within this restricted sample (Figure A.2) and within the entire sample (Figure A.3). None of these specifications change my conclusions.

**IMPLICATIONS**

The results of the field experiment at PickInc show that individual performance pay was superior to group performance pay when it came to boosting productivity and that these effects were not constant across ability levels.
The first finding is in line with existing theory, as PickInc FC’s are an environment low in interdependence. Furthermore, interviews with pickers who received the group incentive indicated that feelings that freeriding was occurring were an obstacle that was hard to overcome.

The second finding is more interesting given existing theory. Against Lazear’s (2000) classic study, the results suggest that individual performance pay may improve motivation among low performers more than among high performers. One reason for this could be that the individual incentive is affecting workers of different ability levels in different ways. For the less able, the incentive may make more salient the minimum standards above which all pickers—regardless of treatment status—are supposed to maintain their performance. Although terminations for poor performance are rare, pickers are expected to pick at a rate near 120 LPH (though this varies by FC). In the participating FCs, this was the rate exhibited by the median picker in the pre-incentive period. For the more able, the incentive may be working as anticipated. It is the picker in the middle who apparently does not see enough benefit in the incentive to work harder. This may be because the cost of increasing their effort does not match up against the amount of money they stand to gain. Group incentives, in this setting, relate to ability as expected.

These results underscore that when weighing performance pay options, firms must consider their ultimate goals in light of the ability distribution in their workforce. Why would a firm choose an individual incentive over a group one, or vice versa? It depends on whether the goal is to bring up top or bottom performers. In settings where interdependence is low and the top is already performing well, the firm may be able to afford losing some of this group’s effort to improve performance at the bottom. A group incentive could work well. Where even the top is underperforming, an individual incentive could help push this group over while also bringing up
weak performers. Those in the middle may require supplementary incentives, perhaps based on subjective rewards. The appropriate incentive may also depend on relative group size. In my setting, performance levels were normally distributed. But in an organization with multiple peaks in ability, it may be best to choose the incentive based on where the biggest gain could be gotten.

The existence of differential responses to incentives in-and-of-itself suggests that optimal performance pay programs need to be tailored to worker preferences. Rather than assuming there is one best policy, a new and improved approach may be to give workers the ability to select from a menu of equal value incentives. There are administrative costs with this option, and it may be infeasible when certain options cannot co-exist, like group incentives for some and individual incentives for others. At the same time, when there are other dimensions over which an incentive can vary, allowing workers to choose their preferred option could maximize performance gains.

LIMITATIONS

This study brings with it several reasons for interpreting its results with caution.

One issue is that the incentives evaluated yielded only limited success. Under the individual incentive, performance improved, but few workers actually took home any reward. This may be because the minimum standard was too high. What is more likely is that other eligibility components, namely attendance, made the increased effort required for payout not worth it in the eyes of workers of middle ability. This group may be evaluating the incentive, then, not in terms of effort but in terms of time: it could more valuable for them to get to work
late than to receive additional income. Consequently, future research should examine additional tradeoffs that employees consider when deciding whether to respond to incentives.

Another is that this study was of limited duration. While there was a short term gain in productivity, at least under the individual incentive, this effect could wane over time, and it could wane for particular ability levels more than others. In fact, managers maintained the individual incentive in the treated FC after the experiment concluded but eventually terminated it due to changes in the business environment. (This outcome will be discussed in a future paper.) Changes in incentives’ effects over time speak to larger issues surrounding administrative challenges, organizational politics, and preferences, which make performance pay programs hard to upkeep. Duration, then, should be important subject of analysis in studies of performance pay.

Finally, there are issues with the design of this study that call for further examination and future research. One concern is that the results were driven by the increased attention workers received, or what is known as a Hawthorne Effect. This concern should be alleviated by the fact that performance is publicly displayed and evaluated daily in picking departments. Hence, PickInc workers are accustomed to evaluation. That said, it is possible, as discussed above, that the positive response of the less able may be because the incentive made performance standards more salient to them. Organizations can still view this outcome as plus. A second concern is that participating FCs were not comparable. It is true that several of their characteristics, like share of full-time and part-time workers, differed. Yet the DD estimation strategy used here, which relies on repeated measures, estimates the incentive effect within individuals rather than simply between workers in the treatment and control FCs. Where resources permit, future research could include a larger number of teams or organizations to see if the results change.
CONCLUSION

This paper shows that, indeed, not everyone is moved by the same motivations at work. Ability shapes how employees respond to performance pay. For researchers studying incentives, there is thus value in moving their analyses away from a focus on average outcomes and toward heterogeneous treatment effects. When it comes to ability in particular, future work should quantify this source of variation relative to other sources of differential responses to incentives, like gender or personality, and it should examine how it changes under different forms of performance pay. Theories that are fine-tuned by research looking at variability can then inform practice at a wider array of organizations. From this paper, practitioners can already take away the conclusion that adopting incentives that complement different aspects of the job and the jobholders is a better strategy than assuming there is one type of incentive that will fit all.
References


Imbens, Guido W., & Wooldridge, Jeffrey M. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5–86.


### Table A1: DD Regression: Effect of Incentives on Productivity, Full Sample

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>9.271***</td>
<td>9.071***</td>
<td>5.844***</td>
<td>5.814***</td>
<td>6.091***</td>
<td>5.927**</td>
</tr>
<tr>
<td></td>
<td>(2.361)</td>
<td>(2.386)</td>
<td>(1.575)</td>
<td>(1.594)</td>
<td>(1.587)</td>
<td>(2.257)</td>
</tr>
<tr>
<td>Group</td>
<td>-5.103</td>
<td>-4.995</td>
<td>-7.319**</td>
<td>-7.365**</td>
<td>-7.098**</td>
<td>-0.767</td>
</tr>
<tr>
<td></td>
<td>(2.884)</td>
<td>(2.875)</td>
<td>(2.326)</td>
<td>(2.351)</td>
<td>(2.351)</td>
<td>(2.082)</td>
</tr>
<tr>
<td>Department Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.001</td>
<td>-0.002***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0005)</td>
<td>(0.0004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual fixed effects? No | No | Yes | Yes | Yes | Yes | Yes
Time fixed effects? No | Yes | No | Yes | Yes | Yes | Yes
Building-specific time trends? No | No | No | No | Yes
Observations 16,288 | 16,288 | 16,288 | 16,288 | 16,288 | 16,288

Note: *p<0.05; **p<0.01; ***p<0.001
DV is daily lines per hour (LPH). Standard errors clustered by individual in parentheses.

### Table A2: DD Regression: Effect of Incentives on Productivity, No Exits or New Hires

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(2.299)</td>
<td>(2.341)</td>
<td>(1.577)</td>
<td>(1.711)</td>
<td>(1.702)</td>
<td>(2.178)</td>
</tr>
<tr>
<td>Group</td>
<td>-4.092</td>
<td>-4.040</td>
<td>-6.380**</td>
<td>-6.378*</td>
<td>-6.275*</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>(2.792)</td>
<td>(2.816)</td>
<td>(2.328)</td>
<td>(2.526)</td>
<td>(2.525)</td>
<td>(2.019)</td>
</tr>
<tr>
<td>Department Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.0003</td>
<td>-0.001*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.0004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual fixed effects? No | No | Yes | Yes | Yes | Yes | Yes
Time fixed effects? No | Yes | No | Yes | Yes | Yes | Yes
Building-specific time trends? No | No | No | No | Yes
Observations 12,985 | 12,985 | 12,985 | 12,985 | 12,985 | 12,985

Note: *p<0.05; **p<0.01; ***p<0.001
DV is daily lines per hour (LPH). Standard errors clustered by individual in parentheses.
Figure A1: Nonlinear Marginal Effect of Incentives on Productivity Moderated by Ability, Rankings Based on Full Sample

Figure A1a: Without building specific time trends.

Figure A1b: With building specific time trends.
Figure A2: Nonlinear Marginal Effect of Incentives on Productivity Moderated by Ability, No New Hires or Exits During Incentive

Figure A2a: Without building specific time trends.

Figure A2b: With building specific time trends.
Figure A3: Nonlinear Marginal Effect of Incentives on Productivity Moderated by Ability, No New Hires or Exits During Incentive, Rankings Based on Full Sample

Figure A3a: Without building specific time trends.

Figure A3b: With building specific time trends.