Employer Expectations, Peer Effects and Productivity: Evidence from a Series of Field Experiments

John J. Horton
Harvard University*

August 27, 2011

Abstract

In five field experiments, workers labeled photographs and evaluated their peers' performance at the same task. Evaluating high-output work made workers more productive, with stronger effects for higher-productivity workers. Even very explicit employer instructions were unable to stamp out these productivity peer effects. In their evaluations, workers punished workers who demonstrated low effort but low output alone was not sufficient to trigger punishment. Willingness to punish was strongly correlated with own productivity, yet this relationship was experimentally mutable, with productivity-reducing treatments also reducing punishment.

JEL J01, J24, J3

*Email: john.joseph.horton@gmail.com. Thanks to the NSF-IGERT Multidisciplinary Program in Inequality & Social Policy (Grant No. 0333403), the University of Notre Dame and the John Templeton Foundation's Science of Generosity Initiative for generous financial support. Thanks to the Centre for Economic Performance at the London School of Economics and Political Science for hosting me while I worked on this project. Thanks to Richard Zeckhauser, Robin Horton, Heidi Yerkes, David Yerkes, Justin Keenan, Larry Katz, Nicholas Christakis, Rob Miller, Malcolm-Wiley Floyd, Renata Lemos, Lydia Chilton and seminar participants at the Wharton Advances with Field Experiments conference for helpful comments and suggestions. Thanks to John Comeau and Alex Breinen for research assistance. All datasets, code and auxiliary regression results are currently or will be available at the author's website: http://sites.google.com/site/johnjosephhorton/.

1 Introduction

In economic theories of employment, firms shape the incentives faced by their workers to obtain compliance with the firm's wishes. For example, a firm might obtain high effort by offering explicit incentive contracts (Holmstrom, 1982), relational contracts (Levin, 2003) or efficiency wages (Katz, 1986). In each of these theories, the focus is squarely on the relationship between the firm's management and the individual employee. A worker's co-workers or "peers," if they matter at all, are relevant only to the extent that they influence the incentives offered by the firm (e.g., by influencing payoffs in a relative performance scheme). However, recent empirical research has highlighted the direct effects that co-workers can have on each other's productivity without intermediation by the firm.[1]

Although peers are important in the workplace, the precise reasons why they matter are not clear, at least in part because there are so many possibilities. Peers might offer instruction, threaten punishment, promise rewards, spur competition or convey firm-specific norms and employer expectations. Even precisely detecting and measuring peer effects is challenging, because group or network formations are often endogenous. Even when formation is exogenous, common shocks and the reflection problem complicate the interpretation of any detected effects (Manski, 1993).

The purpose of this paper is to credibly identify peer effects in a real work setting and to distinguish among mechanisms that could explain these peer effects. In particular, we focus on elucidating the role that peers play in conveying employer expectations. We conducted five sequential experiments in which we hired workers to come up with labels for photographic images.[2] Before agreeing to participate, would-be workers read a description of the task, learned the amount they could earn and viewed a work sample. The work sample was just a "screen shot" of the image-labeling interface as it would look after a worker completed the task.[3] A high-output work sample would show many generated labels, and a lower-output work sample would show fewer labels. In all but the first experiment, we had workers evaluate other workers by inspecting the evaluated workers output.

[1] See, for example, Bandiera et al. (2010), Mas and Moretti (2009), Falk and Ichino (2006), and Guryan et al. (2009).

[2] For example, a photograph of a breakfast scene might generate the labels "juice, toast, cereal."

[3] Because subjects were not informed they were participating in experiments, the experiments were "natural" field experiments in the Harrison and List (2004) taxonomy.

1.1 Overview of experiments and findings

Experiment A laid the groundwork for the follow-on experiments by establishing that (1) workers regard producing output as personally costly and that (2) the output level shown in work samples conveys employer expectations. In Experiment B, workers readily punished peers who demonstrated low output at the image-labeling task. The experiment also demonstrated that productivity and willingness to punish were positively correlated. In Experiment C, the relationship between productivity and punishment was shown to be causal—experimental manipulations that reduced productivity also reduced a worker's willingness to punish. Experiment D demonstrated that evaluating high-output work raised the evaluating worker's subsequent productivity, with effects stronger for workers with higher baseline productivity. Finally, Experiment E showed that peers can still influence a worker's productivity even in the presence of clear, explicit statements of employer expectations. Table 1 summarizes the experimental designs, the research questions and the results for each of the experiments.

1.2 Interpretation

Conditional upon a number of caveats we will discuss in depth, it appears that social learning from peers about employer expectations is an important factor in explaining the pattern of results across the experiments. We find that workers do learn from peers about employer expectations and act in predictable ways in response to this information. Workers are sensitive to peer choices even when employer standards are fully revealed, though perhaps this is because they expect to be evaluated by peers rather than solely by employers. We find that workers are willing to punish low effort, but do not appear to be general-purpose enforcers of employer expectations.

2 Conceptual framework

We are mainly concerned here with how workers learn about employer expectations and how this learning affects productivity. Our focus on "learning" presupposes that employees have something to learn—i.e., that employer expectations are imperfectly known to workers (at least initially). This assumption has theoretical justification: in some models of the firm, the difficulty of writing complete employment contracts (and thus fully specifying expectations) is what drives the firm to use employees rather than meeting its labor needs via outsourcing (Gibbons, 2005).

Rather than write out complete employment contracts, firms integrate or "on board" new employees by providing them instructional materials, by putting them through formal training and by having them interact with and observe more senior employees. By observing the output choices of these existing workers, the new employee can infer what the observed employee believes to be acceptable. It is this peer-driven learning process that we try to re-create in our experimental setting. To do this, our initial instructions were deliberately ambiguous. For example, although we stated how much we would pay, we did not specify output minimums, maximums or bonus criteria, though we did in all cases provide a work sample.[4]

2.1 Social learning does not necessarily promote homogeneity

As new workers integrate into a firm, we expect them to become more "like" more senior workers by learning and using firm-specific jargon and following the firm's idiosyncratic procedures. This kind of acculturation is mainly about coordination—new workers are more or less indifferent to the actual coordinating solution, as making a choice to implement the choice is basically costless. This stands in sharp contrast to choices that have personal costs and that will depend on worker-specific attributes. For example, learning that an employer has either high or low productivity expectations will have a consequential effects on a worker's output choices and will be mediated by a worker's own production costs. This kind of learning—and the action that follows from this learning—can make a new employee *less* like the old employee that they learn from.

We can imagine a scenario in which a worker observes an older, established worker producing very high output relative to what the new worker thought would be sufficient. The new worker could try to raise productivity to match their updated belief about expectations, or the new worker might quit after deciding that meeting the new standard is not worth the cost. To illustrate this more formally, consider an employment scenario in which employers have a standard Y_i where $Y_i \in \{Y_H, Y_L\}$, with $Pr(Y = Y_H) = \theta$ and $Y_L < Y_H$. An employer's type (i.e., their standard) is private information that the employer cannot communicate to workers, but θ is common knowledge. An employer's value from an employment relationship is $v_i(y) = 1 \cdot \{y \geq Y_i\}$, where y is the worker's choice of output level. Workers can choose $y \in \{0, y_L, y_H\}$, with $y_L \geq Y_L$, $y_H \geq Y_H$ but $y_L < Y_H$ and $0 < Y_L$.

[4]Ironically, this "ambiguity" was entirely normal by the apparent standards of the market.

Workers have a type t that is distributed on the interval $[0, 1]$. This type affects their realized costs of output, which can be 0, tc_L or tc_H for y choices of 0, y_L and y_H, respectively. An employer "rejects" (and does not pay for) work if it would get no value in accepting it, i.e., if $v_i(y) = 0$. The worker's utility function is $u(y) = 1\{y \geq Y_i\} - tc(y)$. The worker's expected utility as a function of output is:

$$\mathbf{E}[u(y)] = \begin{cases} 0 & : y = 0 \\ 1 - \theta - tc_L & : y = y_L \\ 1 - tc_H & : y = y_H \end{cases}$$

We can partition the distribution of workers into three intervals: $[0, t_{HL}]$, $[t_{HL}, t_{0L}]$, $[t_{0L}, 1]$. Let t_{HL} be the type of the worker indifferent between producing y_H and y_L. For this worker, $1 - t_{HL}c_H = 1 - \theta - t_{HL}c_L$ which implies that $t_{HL} = \frac{\theta}{c_H - c_L}$. Let t_{0L} be the type of the worker indifferent between producing y_L and not producing at all. For this worker, $0 = 1 - \theta - t_{0L}c_L$, which implies that $t_{0L} = \frac{1-\theta}{c_L}$.

Suppose that a worker is exposed to the output of another worker, and that this work is informative about the employer's expectations (i.e., whether the employer is of Y_H or Y_L type). Suppose the interaction causes the worker to update beliefs to $\theta' > \theta$. We can see that $\delta t_{HL}/\delta\theta = \frac{1}{c_H - c_L}$, meaning that the $[0, t_{HL}]$ interval expands and at least some workers who previously would have produced only y_L increase output to y_H. However, $\delta t_{0L}/\delta\theta = -1$, increasing the size of the $[t_{0L}, 1]$ interval, meaning that more workers now choose $y = 0$.

In the model above, peer effects were not only heterogeneous, but actually differed in sign for different parts of the cost (t) distribution. This example illustrates why the common practice of assuming linear-in-means peer effects for empirical work can be potentially misleading.[5] This fact is useful for interpreting the results that follow.

2.1.1 Evaluation and social preference

In an important feature of experiments B through E, is that workers evaluate other workers. They are asked to split a bonus with the evaluated worker and make a recommendation as to whether or not we should approve the work. We will present the results after describing the experiment, but it is helpful to briefly discuss the choices that evaluating workers make in terms of the model above.

Let $y_i(t)$ be a worker's own output when they spend t unit of time on the task

[5] A recent cautionary tale about making too strong assumptions about the nature of peer effects comes from Carrell et al. (2010), whose experimental estimates of peer effects were opposite in sign to the estimates derived from natural variation.

and let their costs be $c_{self} = \gamma_i t$. They evaluate a fellow worker whose costs they infer to be c_{other}. We will assume that workers make a kind of attribution error and incorrectly believe that all workers have the same productivity and opportunity cost, so that differences in output map directly to differences in time spent, and hence differences in realized costs.

The evaluators are asked to split a bonus B with this other worker. For simplicity, we will assume a highly eqalitarian utility function which is maximized when payoffs are equal between themselves and the other worker. If they believe that the other worker will get their work approved by the employer, they want to transfer an amount x such that $1 - c_{other} + x = 1 - c_{self} + B - x$ which implies that $x^* = B/2 - \Delta c/2$, where $\Delta c = c_{self} - c_{other}$. Workers give equal splits of the bonus to workers who they feel have similar costs, reward workers who have high realized costs (and thus high output) and punish workers who have low costs (low output).

2.2 Related work of peer effects in employment settings

Several recent papers examine the effects of peers on workplace productivity. Perhaps the most illuminating observational evidence comes from Mas and Moretti (2009), who found that less-productive grocery clerks exhibited greater productivity when working near highly productive clerks, but only when they were in the direct view of the highly productive clerks. This finding suggests that the threat of punishment might partially explain workplace peer effects. There is much laboratory literature supporting this punishment-as-peer-effect view, with several studies showing that workers will readily bear costs and altruistically punish peers who free-ride in public goods games (Fehr and Gachter, 2002, 2000). This "strong reciprocity" (Carpenter et al., 2009) is a powerful peer effect, and although firms are not perfectly analogous to public goods games, the notion of worker-enforced productivity norms offers a very general potential solution to the incentive problem of team production.

Guryan et al. (2009) also use evidence from a real workplace, albeit an unusual one: they exploited the random assignment of professional golfers to tournament foursomes to estimate the effects of each player's peers on the player's own performance. In contrast to Mas and Moretti, Guryan et al. found no evidence of peer effects, providing a useful corrective to hasty or overly broad generalizations. However, professional golf tournaments are unusual work environments, in that two common channels for peer effects are foreclosed: professional golfers know what constitutes good performance and they are unlikely to raise their quality of play in the "shadow" of

punishment that might be meted out for non-compliance with productivity norms. In marked contrast with the Mas and Moretti setting, shirking by a professional golfer imposes a positive externality on "co-workers."

Using a field experiment, Falk and Ichino (2006) found that workers stuffing envelopes in pairs had less variation in their output levels than synthetically paired workers constructed from an experimental group whose members worked alone. Though they could not estimate the direction of peer effects (i.e., whether low productivity affects or is affected by high productivity, or some combination of effects), their analysis of the output distribution led them to conclude that it was more likely that less-productive workers were made more productive by working in pairs.

The differences among the Guryan et al. setting, the Falk and Ichino setting and the Mas and Moretti setting—and the resultant differences in findings—motivate the present study, which has an unusual but highly controllable work context that preserves some of the common features of work environments, including uncertainty about norms, costly effort and a task unlikely to inspire much intrinsic motivation. Unlike the Mas and Moretti setting, however, there are no overt free-riding externalities (in the check-out line, slacking by one clerk increases the workload of other clerks). The absence of direct negative externalities is important, as evidence from such a setting can provide some sense of how general punishment might operate in the workplace.

3 Methods and Materials

The experiments were conducted online, using Amazon's Mechanical Turk (MTurk), an online labor market in which workers complete small tasks for payment. MTurk is one of several online labor markets that have emerged in recent years (Frei, 2009). Researchers in a number of disciplines have begun using these markets for experimentation.[6] Some examples in economics include Mason and Watts (2009), Barankay (2010), Chandler and Kapelner (2010) and Horton and Chilton (2010). Online experiments offer advantages yet can be harder to control than conventional laboratory experiments Horton et al. (2010).

[6]For an overview of online labor markets, see Horton (2010).

Table 1: Details of the Experiments

Exp.	Question	Set-up	Treatment	Control	Result
A	Can employers convey productivity expectations?	Subjects viewed an employer-provided work sample, then chose how many labels to produce (if any). Work samples differed by experimental group.	HIGH: Subjects viewed high-output work sample (many labels)	LOW: Subjects viewed low-output work sample (few labels)	HIGH increased labor supply on intensive margin, but decreased it on extensive margin
B	Do workers punish workers that exhibit low productivity?	Subjects viewed an employer-provided work sample, then chose how many labels to produce. Subjects then evaluated another worker's work product.	GOOD: Subjects evaluated a high-output work sample	BAD: Subjects evaluated a low-output work sample	GOOD increased approval recommendations and bonus amounts. Highly productive workers punished more with their evaluations.
C	Is the relationship between own-productivity and punishment causal?	Subjects viewed an employer-provided work sample, then chose how many labels to produce. Subjects then evaluated another worker's work product.	CURB: Subjects received a notice after two labels saying that 3 labels was probably enough output	NONE: Subjects received no notice.	Greatly reduced output in CURB; those in CURB more likely to recommend approval and grant larger bonuses
D	Does exposure to low-output work affect a worker's productivity?	Subjects viewed an employer-provided work sample, then chose how many labels to produce. Subjects then evaluated another worker's work product. Then they labeled a second image.	GOOD: Subjects evaluated a high-output work sample	BAD: Subjects evaluated a low-output work sample	GOOD raised output on second task; effects were stronger for more productive subjects (measured by first task output).
E	Are workers susceptible to peer effects in presence of strongly-stated employer expectations? Do they punish high-effort but non-complying work?	Subjects viewed an employer-provided work sample with 2 labels, then chose how many labels to produce. Subjects were told that 2 and only 2 labels should be produced. Subjects then evaluated another worker's work product, then labeled a second image.	OVER: Subjects evaluated a worker producing too many labels	OK: Subjects evaluated a worker producing the required number of labels	OVER increased subsequent output beyond ceiling, but did not cause more punishment.

3.1 Amazon Mechanical Turk

The tasks posted on MTurk are called "Human Intelligence Tasks" (HITs). HITs vary, but most are small, simple tasks that are difficult for machines but easy for people, such as transcribing audio clips, classifying and tagging images, reviewing documents and checking websites for pornographic content. The creators of HITs are called "requesters." The requester constructs the user interface for their HITs and sets the conditions for payment, worker qualifications and timing (e.g., how long a worker can work on a task). To join, a would-be worker must have a bank account and create a profile with Amazon.[7]

Workers can have only one account, and Amazon uses several technical and legal means to enforce this restriction. Workers can observe the collection of HITs available to them and, in most cases, view a sample of the required work. They can work on any task for which they are qualified and can begin work immediately after accepting a HIT.

Once a worker completes a HIT, the work product is submitted to the requester for review. Requesters may "reject" work, in which case the worker is not paid. The ability of requesters to reject work creates consequences for providing work that does not meet employer expectations. Requesters may also elect to pay bonuses, which makes it easy to tailor payments to individual workers based on their performance within a nominally piece-rate HIT.

3.2 Task and interface

Because computers have trouble recognizing objects in photographs, image labeling is a common "human computation" task (von Ahn and Dabbish, 2004; Huang et al., 2010) that appears frequently in the MTurk market. To actually label images, workers in our experiment used a simple interface we developed, shown in Figure 1. To add a label, workers simply clicked a button labeled "Add a label" positioned below the image. [8] Clicking the button caused a new blank text field to be added to the survey. Workers could add as many labels as they wanted. When they were finished, they clicked a button labeled "Submit labels."

[7] MTurk workers appear to be split approximately evenly between the US and India. Horton (forthcoming) finds workers generally view employers online as having the same level of trustworthiness as offline, traditional employers. For the demographics of the MTurk population, see Ipeirotis (2010).

[8] The images themselves were selected from the photo sharing site Flickr. The images each had a Creative Commons license and were chosen because they were conducive to labeling (e.g., photos depicting elaborate meals with many easily recognizable food types).

As we discussed in the conceptual framework (Section 2), employer expectations for a task can be complex. Even in the simple image-labeling setting, employers could have requirements about both label quantity and label quality, depending on their purposes.[9] In this experiment, we focus only on the quantity of labels workers produced for an assigned image, not the quality of the labels themselves. Our focus on quantity is partly justified by the fact that the merit of any given label is determined by whether the proposed item either appears or does not appear in an image, making the correctness of any particular label fairly objective. It is still possible to imagine an item being more or less central, but we maintain the assumption that restricting attention to quantity alone can still give insights into employee decision-making and peer-effects.

In several of the experiments, workers evaluated the work product of another worker. This evaluation consisted of (1) viewing the image that the evaluated worker was assigned (2) viewing the labels that the evaluated worker provided and (3) deciding whether to recommend rejection and how to split a bonus between themselves and the evaluated worker. It is likely that workers viewed the evaluation and bonus scheme as unextraordinary—it is very common to have workers evaluate the work of other MTurk workers as a way to "bootstrap" quality and bonuses are often used to incentivize good performance.

In each of the five experiments, subjects answered a short demographic survey before beginning work. Subjects were asked their gender, their country and whether they use MTurk primarily in order to make money, learn new skills or have fun. There were small differences in the reported covariates across experiments, most likely due to differences in when experiments were launched. Although one might think the survey would raise suspicions that the task was an experiment, we view this as unlikely:[10] asking workers for basic demographic information is fairly common in the market, as requesters frequently use location and formal qualifications to screen workers.

4 Experiment A: Perceived employer expectations

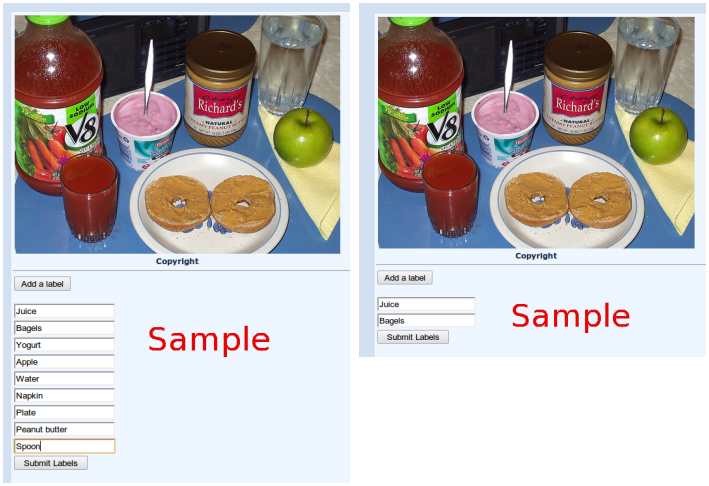
Experiment A tested whether a firm in this market can influence worker productivity by manipulating worker perceptions of employer expectations. Would-be participants were assigned to one of two experimental groups: $HIGH$, in which the work sample

[9] One really good label might be needed for a magazine caption, or a whole collection of mediocre labels might be desired if the the point is return a "loose" set of images in response to a search query.

[10] The IRB approval for these experiments did not require notification that the work was part of a research project.

Figure 1: Work samples shown to workers prior to task acceptance in Experiment A.

(a) $HIGH$ $(y = 9)$ (b) LOW $(y = 2)$



showed 9 labels, and LOW, in which the work sample showed 2 labels. The motivation for showing the work sample was to quickly convey employer expectations. Figure 1 shows the two work samples. After viewing the work sample, workers chose to participate and label an image or exit. Table 2 shows the summary statistics for the experiment and includes details on the sample size, survey results and payment. In this experiment and all subsequent experiments, subjects were assigned to groups by stratifying on arrival time (e.g., subject 1 was assigned to $HIGH$, subject 2 to LOW, subject 3 to $HIGH$ and so on).

4.1 Results

Subjects in $HIGH$ produced absolutely more output. The main results from the experiment are displayed in Figure 2, which shows histograms of output for each experimental group. Mean output is indicated via a vertical line in each panel. We can see that many subjects in $HIGH$ produced more than 12 labels, but only 1 subject in LOW produced more than 12 labels. However, the greater productivity in $HIGH$ came at a cost: a larger fraction of subjects in $HIGH$ elected not to produce any output and quit.

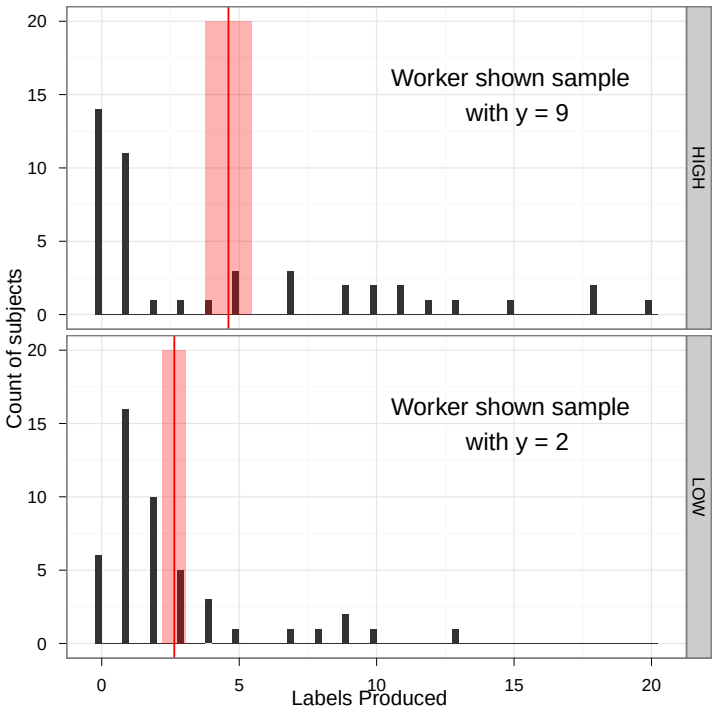


Figure 2: Histogram of labels produced by experimental group in Experiment A. The solid vertical line indicates the mean, while the shaded band around that line is 2 standard errors wide. Subjects in the $HIGH$ group were shown a work sample consisting of 9 labels prior to performing, while subjects in LOW were shown a work sample with only 2 labels. The right edge of each bar intersects the x-axis at the corresponding member of the support: i.e., the largest output choice in the $HIGH$ panel is $y = 20$. This plot and all plots in the document were made using the open-source R package ggplot2 (Wickham, 2008).

Table 2: Experiment A summary statistics ($n = 93$)

Administrative						
Launch: Fri Apr 09 21:10:31 GMT 2010						
Finish: Sun Apr 11 10:04:40 GMT 2010						
Survey	%					
male	58.1					
from India	48.4					
from US	37.6					
motivated by money	76.3					
Treatment Assignment						
$HIGH = 1$	49.5					
Output	Min	.25	Med.	Mean	.75	Max
Labels produced (y)						
in HIGH	0	0	1	4.609	8.5	20
in LOW	0	1	2	2.638	3	13
Entry ($y > 0$)						
in HIGH	0	0	1	0.6957	1	1
in LOW	0	1	1	0.8723	1	1
Time spent on task (seconds)						
in HIGH	9.157	53.27	98.13	172.9	225.4	715.5
in LOW	9.563	28.61	54.9	94.85	110.4	904.3

The key result from Experiment A can be seen in the differences in group means in the "Labels produced" and the "Entry" rows.

4.1.1 High employer expectations reduced labor supply on the extensive margin

Subjects assigned to $HIGH$ were significantly less likely to accept the task compared to subjects in LOW. When we regress an indicator for any output at all on the treatment indicator, we have:[11]

$$1\{y > 0\} = \underbrace{-0.177}_{[0.085]} \cdot HIGH + \underbrace{0.872}_{[0.050]} \tag{1}$$

with $n = 93$ and $R^2 = 0.05$.

[11] Standard errors are robust and shown under the coefficient.

4.1.2 High employer expectations increased labor supply on the intensive margin

Despite the much greater number of subjects in $HIGH$ who chose not to participate (and thus provided 0 labels), output was unconditionally higher in $HIGH$ than in LOW. Even with the non-participants included as $y = 0$ observations, subjects in $HIGH$ produced, on average, roughly 2 more labels per person:

$$y = \underbrace{1.970}_{[0.956]} \cdot HIGH + \underbrace{2.638}_{[0.430]} \tag{2}$$

with $n = 93$ and $R^2 = 0.05$.

4.2 Discussion

Experiment A suggests that workers use the work sample to infer how much work will be required to meet the employer's expectations and thus obtain payment. Given that buyers can reject submitted work, the labor supply results are consistent with workers viewing label creation as costly. Some of the subjects decided that the costs of meeting the perceived requirements for the $HIGH$ group were too high and chose to exit. Those subjects who stayed worked to the higher perceived standard and completed the task. Because unconditional output rose significantly in $HIGH$, we know that selection alone cannot explain the increase in output: it was some combination of selection and greater output.

The results highlight the trade-off that firms might face when communicating standards to employees. Claiming to have high standards may be counterproductive, depending on the nature of the firm's demand for labor. In our particular image-labeling application, conveying high standards via the highly productive work sample was efficient only if the goal was to minimize the per-label price, but it is easy to imagine scenarios in which this is not the objective. For tasks like image labeling, obtaining a large and diverse pool of workers—each contributing a relatively small amount of output—may be more useful than obtaining lots of output from a small number of workers, in which case the high standards conveyed to $HIGH$ would have been undesirable.

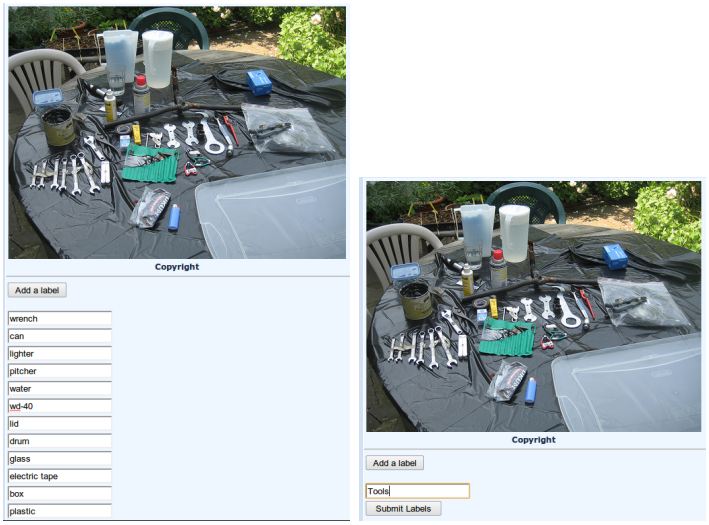
5 Experiment B: Punishment and peer output

Experiment B investigated the conditions under which workers reward or punish their fellow workers on the basis of their co-workers' output. Subjects in the experiment first completed an image-labeling task and then were randomly assigned to either the *GOOD* or the *BAD* experimental group. The *GOOD* subjects inspected the output of a worker from Experiment A that produced 12 unique labels while *BAD* subjects inspected the output of a worker that produced only 1 unique label. The output samples of the evaluated workers are shown in Figure 3. Subjects were then asked to (1) give a recommendation as to whether or not the work inspected should be approved and (2) decide how to split a 9-cent bonus with the evaluated worker. Specifically, subjects were asked, "Should we approve this work?" and had to answer "yes" or "no." Both questions were asked on the same survey page, and subjects could answer them in either order, though the approval question was first on the page. In the regressions that follow, $approve = 1$ indicates that a subject recommended approval. For the contextualized dictator game, subjects were told:

> "We want to determine how good this work is. We would like you to decide, based on your work and the quality of the other work, how to split a 9-cent bonus."

Subjects selected an answer from a list of 9 options of the form "X cents for them, $9-X$ cents for me," with X ranging from 0 to 9 (9 cents was chosen as the endowment to reduce the salience of the focal point 50-50 split). At the end of the experiment, we implemented all choices, with bonuses paid to the evaluating subjects and to the two lucky subjects responsible for the *GOOD* and *BAD* evaluated work samples. In the regressions that follow, the amount transferred to the evaluated subject is represented by *bonus*, with $bonus \in [0, 9]$.

The MTurk job posting for Experiment B was nearly identical to the posting for Experiment A, except that potential subjects were told that they would be evaluating the work of another worker. Before accepting the task, all subjects were shown the *HIGH* work sample used in Experiment A. Because of the additional evaluation work, the participation payment was raised from 30 cents to 40 cents. The requested sample size was also increased to 200. Table 3 reports the summary statistics for Experiment B.



(a) $GOOD$ ($y = 12$) (b) BAD ($y = 1$)

Figure 3: Work evaluated by subjects in Experiment B

Table 3: Experiment B summary statistics ($n = 167$)

Administrative						
Launch: Sun Apr 11 04:36:25 GMT 2010						
Finish: Mon Apr 12 12:39:20 GMT 2010						
Survey	%					
male	58.1					
from India	44.9					
from US	29.3					
motivated by money	75.4					
Treatment Assignment						
$GOOD = 1$	50.9					
Recommended firm approve work						
in GOOD	91.8					
in BAD	50					
Bonus to evaluated worker	Min	.25	Med.	Mean	.75	Max
in GOOD	0	4	5	4.929	6	9
in BAD	0	1	3	3.488	5	9

Notes: Overlap in subjects across experiments was $|A \cap B| = 15$. Subjects in $GOOD$ evaluated work displaying 12 labels, while subjects in BAD evaluated work displaying only 1 label. The key results from the experiments can be seen in the group proportion differences in the "Recommended firm approve work" rows and the group mean differences in the "Bonus to evaluated worker" rows.

5.1 Results

The results from Experiment B can be seen in Figure 4, which contains four histograms, each showing the allocation of the 9-cent bonus. The plots are faceted by experimental group (row) and subject recommendation regarding approval (column). We can see that subjects in *GOOD* were very unlikely to recommend rejection, whereas recommendations for rejection were fairly common among subjects assigned to *BAD*. Few subjects in either group transferred 0 cents, except among those subjects in *BAD* that recommended rejection. For the *BAD*/reject subjects, the modal transfer was 0 cents. Most *GOOD* subjects, as well as a large number of subjects who recommended approval despite being in *BAD*, chose a more or less equitable split of 4 or 5 cents.

One result to note in Figure 4 in the *GOOD*/approval quadrant is how generous this distribution is compared to the usual results of the dictator game. In most laboratory dictator games, transfers of 0% or 50% of the endowment are common, with very few subjects transferring more than 50% (Engel, 2010). Yet a clear majority of subjects in *GOOD* transferred amounts greater than 50% of the endowment. This difference is probably due to the very low stakes used in the experiment.

5.1.1 Workers more likely to advocate no pay for lower-productivity work

Confirming what was evident graphically, subjects in *GOOD* were far more likely to recommend approval. Regressing an indicator for approval on the treatment indicator, we have:

$$approve = \underbrace{0.418}_{[0.064]} \cdot GOOD + \underbrace{0.500}_{[0.056]} \tag{3}$$

with $n = 167$ and $R^2 = 0.21$.

5.1.2 Workers rewarded better work with more generous bonuses

Subjects were more generous to their highly productive peers. Regressing the amount transferred on the treatment indicator, we have:

$$bonus = \underbrace{1.442}_{[0.393]} \cdot GOOD + \underbrace{3.488}_{[0.298]} \tag{4}$$

with $n = 167$ and $R^2 = 0.08$.

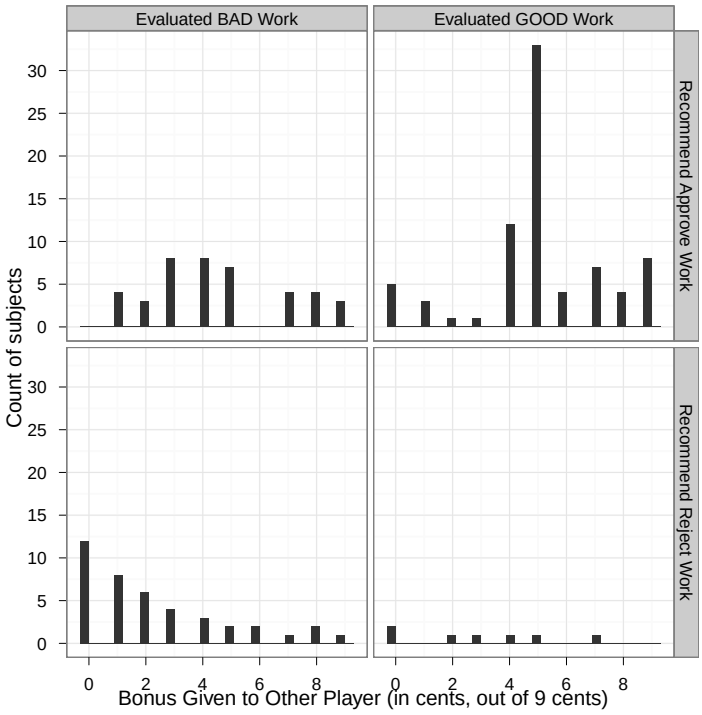


Figure 4: Experiment B, bonus allocation by treatment group and accept/reject recommendation. Subjects in *BAD* evaluated work with 1 generic label, while subjects in *GOOD* evaluated work with 12 specific and appropriate labels.

5.1.3 Highly productive workers less generous and more likely to advocate rejection

There is a strong negative correlation between a subject's own productivity and their generosity in the dictator game:

$$bonus = \underbrace{-0.259}_{[0.066]} \cdot y + \underbrace{5.376}_{[0.365]} \tag{5}$$

with $n = 167$ and $R^2 = 0.07$. The analogous effect also appears strongly in the approval recommendation:

$$approve = \underbrace{-0.055}_{[0.011]} \cdot y + \underbrace{0.960}_{[0.049]} \tag{6}$$

with $n = 167$ and $R^2 = 0.11$.

5.2 Discussion

The output level of the evaluated work had a strong causal effect on measures of both punishment and generosity. Subjects who evaluated low-output work were far more likely to recommend rejection and transfer smaller amounts of money in the dictator game.[12] The simplest explanation may be that workers believe that their evaluations themselves may be spot-checked, and thus it is reasonable for them to adopt whatever they believe would be the principal's view. In other words, they reject poor work because they believe that the employer/requester is likely to believe it is bad. What is less clear is why highly productive workers are more likely to reject work.

There are at least three possible explanations why workers punish their less-productive peers. One possibility is that there are different "types" of workers, and highly productive types (measured by producing high output on the initial image-labeling task) are more likely to punish those they perceive to be less hard-working. Another possibility is that workers idiosyncratically differ in the perception of the prevailing productivity norm, and these differences in norm perception determine both output choices and norm enforcement. Finally, it is possible that workers are inequity-averse (Fehr and Schmidt, 1999), and because they regard productivity as

[12] Greg Little et al. (2010) find that when MTurk workers evaluate the work product of others, they often use readily available metrics such as quantity rather than quality.

costly, they punish low-effort workers and reward high-effort workers as a way to equalize outcomes. This last argument was basically the one sketched out in Section 2.

6 Experiment C: Productivity and punishment

By definition, one cannot experimentally manipulate "innate types." However, if manipulations of productivity cause a change in willingness to punish, then the innate-types hypothesis is untenable. To distinguish inequity aversion from norm enforcement, one would need a way of manipulating a worker's experienced productivity without changing either their perceptions of the prevailing norm or their perceptions of each party's payoffs. Given our imperfect understanding of how workers make decisions about both labor supply and generosity, it is unlikely that the exclusion restriction could be credibly satisfied. Nevertheless, ruling out the innate-types hypothesis is still worthwhile, which is the goal in Experiment C.

To illustrate the challenge of distinguishing among the hypotheses, consider the implications of using the setup of Experiment A to induce changes in labor supply. In Experiment A we were able to change output by altering the work sample shown to workers before they accepted the task. We did not measure follow-on output in a second image-labeling task in Experiment A nor did we have workers evaluate others' work. If we had, at least two problems would have arisen. First, the manipulation would have had an effect on both the intensive margin and the extensive margin. Second, subjects who quit in the first stage would not have recorded their choices in the dictator game or their answer to the accept/reject question.

The fundamental problem is that any intervention that changes worker pre-uptake beliefs about the work required—and hence the labor supply on the extensive margin—is likely to create a missing data problem. For this reason, the intervention used in Experiment C changes worker productivity *after* subjects have already begun working. The design, as well as the failed pilots, will be discussed in detail, but the key point is that the intervention successfully manipulated output on the intensive margin and yet did not cause any across-group differences on the extensive margin (i.e., lead to differences in group composition). However, it likely did affect perceived deservingness or inequity, making it impossible to decide between the two hypotheses related to these concepts.

6.1 Pilot experiments

Prior to running Experiment C, two failed pilot experiments were conducted. In the first pilot, half of the subjects were assigned to work with an interface containing a hidden "bug" that introduced a 1.2-second delay in the software execution after each label was added; those in the control faced no delay. This pilot failed because it did not generate a "first stage" by reducing output: although workers in the "slow" treatment took longer, they did not produce any less output. This outcome is consistent with other findings that workers on MTurk appear to be insensitive to small differences in the time it takes to complete a task (Horton and Chilton, 2010).

In the second pilot, subjects in the "slow" treatment received a pop-up box containing the text "Thank you! Three is probably enough." after they had added a second label but before they added a third label. Although this treatment did have a large effect on worker productivity, other aspects of the design of the experiment generated little variation in generosity. The first problem was that in order to obtain a larger sample, the LOW work sample from Experiment A was used, thereby compressing the productivity distribution (as in Experiment A). The second problem was that workers evaluated work with 3 good labels. As a result, regardless of their treatment assignment, many subjects chose the pseudo-50-50 (i.e., chose either the 4- or 5-cent transfer) split and recommended approval, providing little useful variation in measurements of punishment and generosity.

6.2 Actual experiment

After the two failed pilots, the second pop-up notice experiment was relaunched, albeit with two modifications. The $HIGH$ 9-label work sample from Experiment A served as the sample, and the evaluated work was particularly bad—the evaluated worker provided only 1 generic label for an item-rich photograph. Subjects were assigned to either $NONE$, in which they were given no notice, or $CURB$, in which they received the pop-up notice after completing a second label. Because two-stage least squares would be used in the data analysis, the total sample size was increased to 300. Payment was 30 cents. Table 4 reports the summary statistics for experiment.

6.3 Results

Worker output and acceptance recommendations at the image-labeling task are shown in Figure 5. The x-axis is broken into discrete output ranges, while the y-axis is the

Table 4: Experiment C summary statistics ($n = 273$)

Administrative						
Launch: Sun Apr 18 19:36:44 GMT 2010						
Finish: Wed Apr 21 05:18:02 GMT 2010						
Survey	%					
male	57.1					
from India	35.9					
from US	40.7					
motivated by money	70.7					
Treatment Assignment						
$CURB = 1$	51.3					
Recommended we approve work?						
in CURB	67.1					
in NONE	57.1					
Labels produced (y)	Min	.25	Med.	Mean	.75	Max
in CURB	1	2	3	2.979	4	10
in NONE	1	1	5	6.15	9	26
Bonus to evaluated worker						
in CURB	0	2	4	3.871	5	9
in NONE	0	1	3	3.075	5	9

Notes: Overlap in subjects across experiments was $|C \cap B| = 20$, $|C \cap A \cap \overline{B}| = 20$. Subjects in $CURB$ reduced an output-reducing notification after adding a second label. The key results from the experiment can be seen in the mean differences across groups (in the "Bonus to evaluated worker" rows and the "Labels produced" rows).

number of subjects. Each point indicates the number of subjects in a particular output band, making a particular recommendation (accept or reject). Points are connected to indicate the trends.

For points connected by dotted lines, subjects rejected the work, while for points connected by solid lines, subjects accepted the work. The top panel shows the results for $NONE$, while the bottom shows the results for $CURB$. Several features of the data are readily apparent. First, assignment to $CURB$ dramatically reduced output: a large number of subjects in $NONE$ produced $y \in (5, 10]$ or $y \in (10, 30]$. By comparison, fewer than 10 subjects total in $CURB$ produced this much. Second, subjects choosing high levels of productivity in $NONE$ were far more likely to recommend rejection. Although bonus allocation is not shown in the figure, this same pattern appears: subjects with reduced productivity (those in $CURB$) were more generous in their allocation of the 9 cents.

6.3.1 Workers with reduced output more generous

There is a strong negative correlation between the amount transferred to the other player in the dictator game and a subject's own output in the image-labeling task, as shown by:

$$bonus = \underbrace{-0.207}_{[0.039]} \cdot y + \underbrace{4.418}_{[0.223]} \tag{7}$$

with $n = 273$ and $R^2 = 0.12$. Assignment to $CURB$ had a strong negative effect on worker output, as shown by:

$$y = \underbrace{-3.172}_{[0.490]} \cdot CURB + \underbrace{6.150}_{[0.470]} \tag{8}$$

with $n = 273$ and $R^2 = 0.14$. The F-statistic for the model is 43.982. The two-stage least-squares estimate of the effect of output on transfers in the dictator game is:

$$bonus = \underbrace{-0.251}_{[0.090]} \cdot y + \underbrace{4.619}_{[0.433]} \tag{9}$$

There was no meaningful difference between the least-squares estimate of the effect of productivity on bonuses and the two stage least-squares estimate.

6.3.2 Workers with reduced output less likely to punish

As in Experiment B, highly productive subjects were more likely to recommend that we not approve the evaluated worker's work:

$$approve = \underbrace{-0.042}_{[0.005]} \cdot y + \underbrace{0.811}_{[0.037]} \tag{10}$$

with $n = 273$ and $R^2 = 0.13$. The two-stage least-squares estimate is:

$$approve = \underbrace{-0.032}_{[0.017]} \cdot y + \underbrace{0.765}_{[0.083]} \tag{11}$$

which is not significantly different from the least-squares estimate.

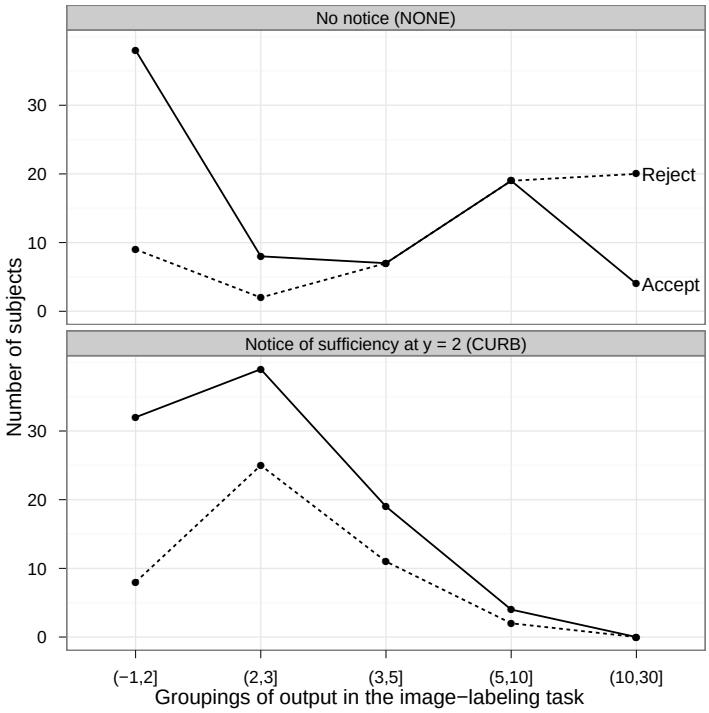


Figure 5: Numbers of subjects recommending either acceptance or rejection at different output levels (indicated by line types), faceted by experimental groups, in Experiment C . Note that a disproportionate number of subjects not receiving the output-curtailing message (subjects in $NONE$) produced relatively high levels of output and that subjects in that high-output band were much more likely to recommend rejection than low-output subjects were.

6.4 Discussion

Experiment C rules out the possibility that productivity and generosity are jointly determined by some innate worker "type": reducing productivity reduced willingness to punish. However, the experiment does not distinguish between the "enforced norms" hypothesis and the "inequity aversion" hypothesis. The problem stems in part from the difficulty in manipulating worker productivity without also altering workers' perception of the relevant norm. It seems possible that future research could disentangle these hypotheses with a suitable experimental design.

Although the results of Experiment C do not favor either hypothesis, other experimental results push the balance of evidence toward the enforced norms hypothesis. First, more complex laboratory games such as those conducted by Charness and Rabin (2002) show that workers trade off a number of competing interests when making dictator-game allocations and that preferences are more nuanced than simple inequity aversion. List and Cherry (2008) make a compelling argument that what we often interpret as "social preferences" in the dictator game is in fact a desire to be seen as complying with some context-specific norm.

7 Experiment D: Peer effects from evaluation

Experiment A showed that exposure to employer-provided work samples affected labor supply, presumably by changing workers' beliefs about the employer's output expectations. Experiment D tested whether work samples from peers that are not held up as examples can still influence productivity. In set-up, Experiment D was similar to Experiment B in that after an initial task, subjects were assigned to one of two groups, $GOOD$ and BAD. In $GOOD$, subjects evaluated a worker who produced 11 labels; in BAD, subjects evaluated a worker who produced only 2 labels. Unlike in Experiment B, however, subjects performed an additional image-labeling task after the evaluation. Table 5 reports the summary statistics for the experiment. The requested sample size was 300 and payment was 40 cents.

7.1 Results

Exposure to the work of a peer strongly affected a subject's subsequent output; output following exposure to highly productive peers was higher than output following exposure to less productive peers. The effect was modulated by productivity on the first task, with workers who were initially more productive being more susceptible

Table 5: Experiment D summary statistics ($n = 275$)

Administrative						
Launch: Fri Apr 23 18:37:56 GMT 2010						
Finish: Wed Apr 28 12:30:17 GMT 2010						
Survey	%					
male	52.4					
from India	35.6					
from US	44					
motivated by money	72.4					
Treatment Assignment						
$GOOD = 1$	48.4					
Labels produced	Min	.25	Med.	Mean	.75	Max
Initial output, before evaluation (y_1)						
in GOOD	1	2	5	4.759	7	14
in BAD	1	2	5	4.768	7	15
Follow-on output, after evaluation (y_2)						
in GOOD	0	4	7	7.368	10	23
in BAD	0	1.25	5	5.014	7	16

Notes: Overlap in subjects across experiments was $|D \cap C| = 26$, $|D \cap (A \cup B) \cap \overline{C}| = 37$. Subjects in $GOOD$ evaluated high-output work, while subjects in BAD evaluated low-output work. The key finding from this experiment was the effect exposure had on subsequent output. We can see that there were no differences in output means pre-exposure ("Initial output, before evaluation" rows) but a large difference after evaluation ("Follow-on output, after evaluation" rows).

to peer effects. Figure 6 is a scatter plots of subsequent output, y_2, against initial output, y_1. The regression line for $GOOD$ is everywhere above the line for BAD and it is steeper.

7.1.1 Exposure to highly productive peers increased productivity

Subjects that evaluated work from highly productive peers produced considerably more labels in the follow-on task:

$$y_2 = \underbrace{0.914}_{[0.071]} \cdot y_1 + \underbrace{2.362}_{[0.373]} \cdot GOOD + \underbrace{0.654}_{[0.398]} \tag{12}$$

with $n = 275$ and $R^2 = 0.49$. In addition to the treatment effect, we can see that initial output was highly correlated with subsequent output. The effect of $GOOD$

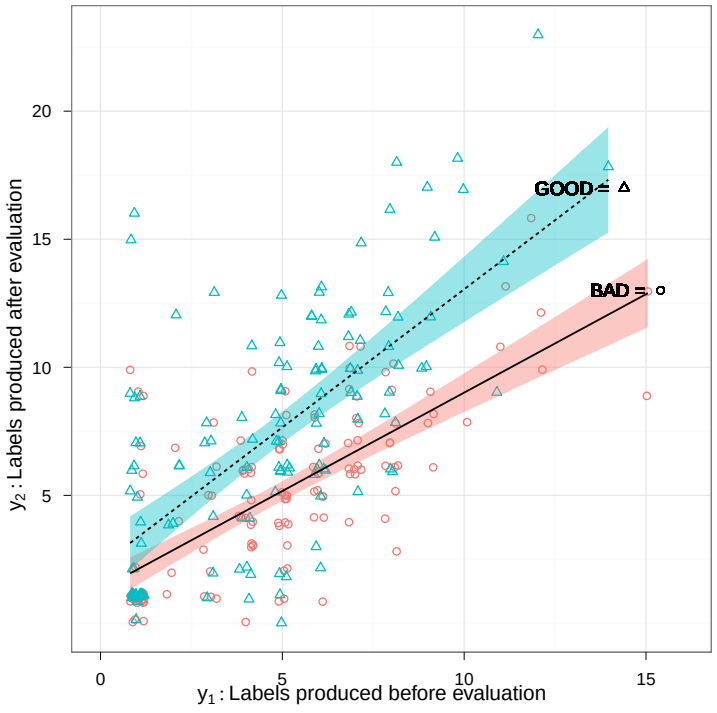


Figure 6: Follow-on output, y_2, versus initial output,y_1, in Experiment D, by group. All subjects did an identical initial task and chose some number of labels to provide (shown on the x-axis). Subjects then evaluated another subject's work that demonstrated either low productivity (BAD) or high productivity ($GOOD$). All output levels are randomly perturbed by $\epsilon \sim U[-.2, .2]$ to prevent over-plotting (which would occur because only integer output levels were possible).

was not, however, constant across the initial output distribution:

$$y_2 = \underbrace{0.771}_{[0.073]} \cdot y_1 + \underbrace{0.317}_{[0.142]} \cdot y_1 GOOD + \underbrace{0.850}_{[0.799]} \cdot GOOD + \underbrace{1.337}_{[0.411]} \tag{13}$$

with $n = 275$ and $R^2 = 0.5$; as y_1 increases, the positive effect of assignment to $GOOD$ on output grows larger.

7.2 Discussion

The size of the peer effects detected in Experiment D strongly depends upon a worker's initial output. The model in Section 2 gave a very general explanation for why peer effects could differ depending upon a worker's initial output. In the context of Experiment D, although being in $GOOD$ and observing highly productive work might still have some effect on beliefs for initially low-productivity workers, these less

productive workers might have fairly inflexible opinions regarding what constitutes acceptable work. Recall that all subjects were exposed to the $HIGH$ work sample from Experiment A prior to accepting the task, and yet some still chose to produce only 1 or 2 labels initially.

Experiment D demonstrated that exposure to peer output affects a worker's own output. Given the pattern of results, it seems likely that workers' output choices for the second task incorporate what they learn from the evaluation. The question remains whether they incorporate this new information because they suspect they will be evaluated by other employees or by their actual employer. Given that they make their output choice immediately after evaluating another worker, a worker *should* be sensitive to the implied standards of other workers (particularly given the results from Experiments B and C, which showed an unequivocal tendency to punish "bad" work). In Experiment E, we investigate this hypothesis by making employer expectations unambiguous and then seeing if exposure to non-complying peers can still influence subsequent output.

8 Experiment E: Peer effects after explicit employer instructions

If peer effects reflect learning about employer standards, then clear, strongly stated production standards should "inoculate" workers from learning-driven peer effects. However, if workers fear punishment by fellow workers or if they have some innate desire to produce as much as peers do, then we should detect peer effects even when standards are clearly communicated.

In Experiment E, these ideas were tested by providing subjects with very explicit instructions about productivity expectations and then exposing subjects to peers. The set-up was almost identical to that of Experiment D, except that workers were told that they should produce 2 and only 2 labels per image. The requirement of 2 labels was stated before workers began the task, and was repeated again with each of the two image-labeling tasks, directly above the image. After performing the initial task, workers were assigned to one of two groups: OK, in which subjects evaluated a work sample showing $y = 2$, and $OVER$, in which subjects evaluated a work sample showing $y = 11$. After evaluating the work, subjects performed an additional image-labeling task. Table 6 shows the summary statistics for the experiment. The requested sample size was 300 and the payment was 40 cents.

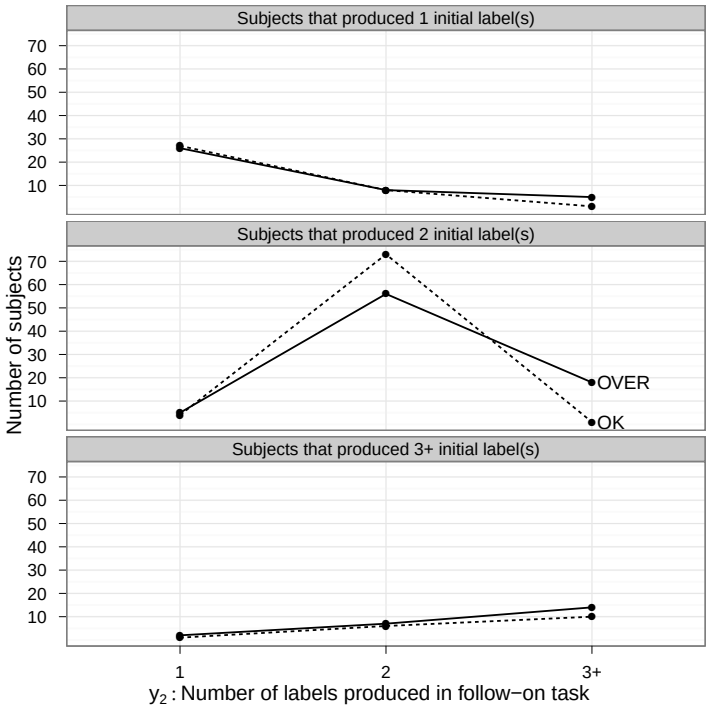


Figure 7: Numbers of workers in different output bands for the follow-on task in Experiment E, faceted by initial output, with experimental groups indicated by line type . Note that for initially complying subjects (middle panel), assignment to $OVER$ increased the relative number of subjects producing additional (and hence non-complying) output in the second task.

Table 6: Experiment E summary statistics ($n = 272$)

Administrative						
Launch: Wed May 19 21:19:42 GMT 2010						
Finish: Sun May 23 15:26:30 GMT 2010						
Survey	%					
male	55.5					
from India	50					
from US	36.8					
motivated by money	77.2					
Treatment Assignment						
$OK = 1$	48.2					
Labels produced	Min	.25	Med.	Mean	.75	Max
Initial output, before evaluation (y_1)						
in OK	1	1	2	2.038	2	8
in OVER	1	1	2	2.085	2	8
Follow-on output, after evaluation (y_2)						
in OK	1	2	2	1.977	2	8
in OVER	1	2	2	2.667	3	14

Notes: Overlap in subjects across experiments was $|E \cap D| = 29$, $|E \cap (A \cup B \cup C) \cap \overline{D}| = 42$. All subjects received explicit instructions to produce 2 labels. After performing one task, subjects were assigned to OK, where they viewed complying work, or $OVER$, where they viewed high-output/non-complying work. The key finding from the experiment can be seen in difference in group means in the "Follow on output, after evaluation" rows.

8.1 Results

In Figure 7, the three vertically stacked panels correspond to three output "bands" that workers could have chosen for the initial task: $y_1 = 1$, $y_1 = 2$ and $y_1 \geq 3$. Within each panel, the line plot shows the number of subjects in each output band for the follow-on task, i.e., $y_2 = 1$, $y_2 = 2$ and $y_2 \geq 3$, broken down by experimental group.

In all three panels and for each group within a panel, the most common output band for y_2 is the corresponding initial output for that panel. In other words, initial output was predictive of follow-on output. We can also see from the middle panel that most subjects chose $y_1 = 2$, suggesting that employer instructions to produce exactly 2 labels were influential.

In the top and bottom panels, there are no differences across experimental groups—the lines are nearly overlapping. The exception is in the middle panel. For those subjects who complied initially, exposure to $OVER$ reduced compliance in the follow-on

task (the dotted line is above the solid line) by increasing output (the dotted line is below the solid line at $y_2 \geq 3$).

8.1.1 Most workers compliant with employer output requests

In both $OVER$ and OK, we can see that $y_1 = 2$ was by far the most common output choice. Reassuringly, Figure 7 shows that the breakdown of y_1 appears to be almost identical across the two treatments. This is expected since subjects were randomized and the experience of the groups did not differ until after the first task was completed. The instruction to produce only two labels appears to have been salient, especially considering that the first stage of this experiment was identical to that of Experiment A, LOW (Figure 2, bottom panel), except that no explicit instructions were given, and in the Experiment A case, $y = 2$ was not the modal choice, as it was in Experiment E.

8.1.2 Language barriers likely prevented full initial compliance

Although not causal, a regression of the compliance indicator on self-reported country suggests that language barriers might have limited compliance, with subjects from India significantly less likely to comply:

$$1\{y_2 = 2\} = \underbrace{-0.228}_{[0.059]} \cdot INDIA + \underbrace{0.691}_{[0.040]} \tag{14}$$

with $n = 272$ and $R^2 = 0.05$.

8.1.3 Evaluating compliant work increased compliance for initially complying workers

Being assigned to OK strongly increased y_2-compliance:

$$1\{y_2 = 2\} = \underbrace{0.161}_{[0.059]} \cdot OK + \underbrace{0.504}_{[0.042]} \tag{15}$$

with $n = 272$ and $R^2 = 0.03$. This effect is driven by subjects in OK who initially complied and then continued to comply on the second image. This is evidenced by the large and significant coefficient on the compliance $\times$ assignment interaction:

$$1\{y_2 = 2\} = \underbrace{0.022}_{[0.083]} \cdot OK + \underbrace{0.205}_{[0.102]} \cdot [OK \times 1\{y_1 = 2\}] + \underbrace{0.467}_{[0.076]} \cdot 1\{y_1 = 2\} + \underbrace{0.242}_{[0.055]} \tag{16}$$

with $n = 272$ and $R^2 = 0.36$. Note that exposure to OK has no effect for originally non-complying workers who chose $y_1 \neq 2$. Also note that initial compliance was strongly predictive of subsequent compliance.

8.1.4 Workers did not punish non-complying work that demonstrated high effort

Assignment to OK had no effect on subjects' accept/reject recommendations:

$$approve = \underbrace{-0.002}_{[0.041]} \cdot OK + \underbrace{0.872}_{[0.028]} \tag{17}$$

with $n = 272$ and $R^2 = 0$. For transfers in the dictator game:

$$bonus = \underbrace{-0.397}_{[0.263]} \cdot OK + \underbrace{5.305}_{[0.182]} \tag{18}$$

with $n = 272$ and $R^2 = 0.01$. Note that the bonus level itself was quite high, and most subjects transferred a more than equitable split. Given the strong positive correlation between subjects' own productivity and self-serving behavior in the dictator game, the large average bonus size is consistent with the fact that most subjects showed low productivity on the initial task (because they were induced to choose $y_1 = 2$). Although the coefficient on OK in the regression above is not significant, it is not a precisely estimated zero, as in the case of approval in Equation 17. Self-serving behavior among subjects assigned to OK is concentrated among highly productive types evaluating the 2-label evaluated work. We can see this by the large negative coefficient on the group/productivity ($y_1 \times OK$) interaction term:

$$bonus = \underbrace{0.720}_{[0.502]} \cdot OK + \underbrace{0.020}_{[0.152]} \cdot y_1 + \underbrace{-0.547}_{[0.204]} \cdot y_1 \times OK + \underbrace{5.264}_{[0.383]} \tag{19}$$

with $n = 272$ and $R^2 = 0.05$.

8.1.5 Exposure to highly productive work increased output even when expectations were explicit

Workers exposed to $OVER$ increased their output compared to those exposed to OK:

$$y_2 = \underbrace{0.658}_{[0.120]} \cdot y_1 + \underbrace{-0.659}_{[0.183]} \cdot OK + \underbrace{1.294}_{[0.299]} \tag{20}$$

with $n = 272$ and $R^2 = 0.25$. However, unlike in Experiment D, this effect was not conditional upon initial output. When the regression above is augmented with an $y_1 \cdot OK$ interaction term (not shown), the coefficient on the interaction is small and insignificant, which is consistent with there being little variation in initial output (i.e., the distribution of y_1 is heavily concentrated at $y_1 = 2$).

8.2 Discussion

It is clear that many workers take explicit requests from employers as informative and worth complying with. Yet a substantial number of workers remained susceptible to peer effects that could, in principle, have led them to have their work rejected for not following explicit instructions. There are several possible explanations.

One possibility is mistake. Workers might believe that they misinterpreted the instructions or that other workers have some inside knowledge. Workers might reasonably believe that we had free-disposal of extra labels and added a few more to provide a margin of safety. However, the requirement was clearly stated at least three times to workers, and many initially complying workers were still pulled upward by highly productive peers.

Another possibility is that workers want to produce an amount comparable to that of their peers, regardless of employer instructions. Given the propensity of peers to punish low effort, matching the output of one's peers is a good idea if there is a chance that one will be evaluated by those peers. Because subjects are asked to evaluate workers, it seems likely that they infer that they will also be evaluated by other workers. Given that other workers are likely to reward or punish based on apparent effort, not necessarily on compliance with the stated standards, above-standard output could be rational even if it is technically non-complying.

9 Conclusion

This paper reports a number of results: (1) workers readily punish low-effort work; (2) workers are susceptible to peer effects, but the effects are conditional upon a worker's productivity; (3) a worker's willingness to punish is mediated by their own productivity; (4) workers punish low effort, not failure to comply with an employer's instructions. Some of these findings contradict or at least complicate results from other workplace settings.

For example, Falk and Ichino found that "bad apples far from damaging good apples seem instead to gain in quality when paired with the latter." Mas and Moretti find a similar result. In the setting examined here, something like the traditional bad apples metaphor applied: it was better to be around good apples than bad apples, and good apples did nothing for the bad apples.

It is perhaps unsurprising that workers punished their low-output peers, as the evaluators likely regarded evaluation to be one of their duties as a (closely monitored) agent of the principal. However, workers were not mindless, general-purpose enforcers of employer expectations: non-complying but high-effort work was not punished. If this distinction generalizes, firms might experience downsides to worker-driven norm enforcement. For example, it may be difficult to get workers to substitute easy, correct procedures for difficult, inefficient procedures. Ironically, the difficulty itself might make an outdated procedure harder to replace, as workers who adopt the easier method might be perceived to be shirking.

One interesting implication is that susceptibility to peer effects, when combined with a causal relationship between own-productivity and willingness to punish, suggest the possibility of a feedback loop, leading to either a virtuous or a vicious cycle. For example, after observing idiosyncratically bad work, workers may lower their own output and punish less in response, in turn reducing other workers' incentives to be highly productive. This whole process can also run in reverse. This may explain why organizational leaders often use the language of contagion to describe morale and why so much of management theory focuses on understanding and influencing organizational culture (Schein, 2004), rather than, for example, trying to write perfectly complete employment contracts.

A natural question for managers is whether they should encourage peers to influence each other, such as by optimally arranging teams to maximize productivity. Here, context seems to matter greatly. Giving workers thicker sticks or juicier carrots to use on their peers may backfire if workers are enforcing inefficient norms. A

long line of research has suggested that workers will enforce norms that are directly counter-productive, particularly if enforcing this norm is in their self interest (e.g., Roy's 1952 work on machine shops). Encouraging norm enforcement when there is uncertainty about what, precisely, will be enforced is risky. Clearly further study is needed.

References

Bandiera, O., I. Barankay, and I. Rasul, "Social incentives in the workplace," *Review of Economic Studies*, 2010, *77* (2), 417–458.

Barankay, Iwan, "Rankings and Social Tournaments: Evidence from a Field Experiment," *Working Paper*, 2010.

Carpenter, J., S. Bowles, H. Gintis, and S.H. Hwang, "Strong reciprocity and team production: Theory and evidence," *Journal of Economic Behavior & Organization*, 2009, *71* (2), 221–232.

Carrell, S., B. Sacerdote, and J. West, "From Natural Variation to Optimal Policy: A Cautionary Tale in How Not to Improve Student Outcomes," *Working paper*, 2010.

Chandler, D. and A. Kapelner, "Breaking monotony with meaning: Motivation in crowdsourcing markets," *University of Chicago mimeo*, 2010.

Charness, G. and M. Rabin, "Understanding social preferences with simple tests," *Quarterly Journal of Economics*, 2002, *117* (3), 817–869.

Engel, C., "Dictator games: A meta study," *Working Paper Series of the Max Planck Institute for Research on Collective Goods*, 2010.

Falk, A. and A. Ichino, "Clean evidence on peer effects," *Journal of Labor Economics*, 2006, *24* (1).

Fehr, E. and K.M. Schmidt, "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

_ **and S. Gachter**, "Cooperation and punishment in public goods experiments," *American Economic Review*, 2000, pp. 980–994.

_ **and** _ , "Altruistic punishment in humans," *Nature*, 2002, *415* (6868), 137–140.

Frei, Brent, "Paid crowdsourcing: Current state & progress toward mainstream business use," *Whitepaper Produced by Smartsheet.com*, 2009.

Gibbons, R., "Four formal (izable) theories of the firm?," *Journal of Economic Behavior & Organization*, 2005, *58* (2), 200–245.

Guryan, J., K. Kroft, and M.J. Notowidigdo, "Peer effects in the workplace: Evidence from random groupings in professional golf tournaments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 34–68.

Harrison, G.W. and J.A. List, "Field experiments," *Journal of Economic Literature*, 2004, *42* (4), 1009–1055.

Holmstrom, B., "Moral hazard in teams," *The Bell Journal of Economics*, 1982, *13* (2), 324–340.

Horton, J.J., "Online Labor Markets," *In Proceedings of 6th Workshop on Internet and Network Economics*, 2010.

_ , "The Condition of the Turking class: Are Online Employers Fair and Honest?," *Economic Letters*, forthcoming.

_ **and L.B. Chilton**, "The labor economics of paid crowdsourcing," in "Proceedings of the 11th ACM Conference on Electronic Commerce" 2010.

_ **, D.G. Rand, and R. J. Zeckhauser**, "The online laboratory: Conducting experiments in a real labor market," *NBER Working Paper w15961*, 2010.

Huang, E., H. Zhang, D.C. Parkes, K.Z. Gajos, and Y. Chen, "Toward automatic task design: A progress report," in "Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)" 2010.

Ipeirotis, P.G., "Demographics of Mechanical Turk," *New York University Working Paper*, 2010.

Katz, L.F., "Efficiency wage theories: A partial evaluation," *NBER macroeconomics annual*, 1986, *1*, 235–276.

Levin, J., "Relational incentive contracts," *American Economic Review*, 2003, *93* (3), 835–857.

List, J.A. and T.L. Cherry, "Examining the role of fairness in high stakes allocation decisions," *Journal of Economic Behavior & Organization*, 2008, *65* (1), 1–8.

Little, G., L.B. Chilton, , M. Goldman, and R. Miller, "Exploring iterative and parallel human computation processes," in "ACM-KDD (HCOMP)" ACM 2010.

Manski, C.F., "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 1993, *60* (3), 531–542.

Mas, A. and E. Moretti, "Peers at work," *American Economic Review*, 2009, *99* (1), 112–145.

Mason, W. and D.J. Watts, "Financial incentives and the 'performance of crowds'," in "Proc. ACM SIGKDD Workshop on Human Computation (HCOMP)" 2009.

Roy, D., "Quota restriction and goldbricking in a machine shop," *American Journal of Sociology*, 1952, *57* (5), 427–442.

Schein, E.H., *Organizational culture and leadership*, Jossey-Bass Inc Pub, 2004.

von Ahn, L. and L. Dabbish, "Labeling images with a computer game," in "Proceedings of the ACM SIGCHI conference on Human factors in computing systems" 2004, pp. 319–326.

Wickham, H., "ggplot2: An implementation of the grammar of graphics," *R package version 0.7, URL: http://CRAN.R-project.org/package=ggplot2*, 2008.