Wielding Peer Effects in Online Production: Evidence from a Series of Field Experiments

John J. Horton*　　Richard J. Zeckhauser
NYU Stern　　Harvard University

October 3, 2018

Abstract

With the rise of online labor markets, remote work, and other forms of computer-mediated production, firms have an increasing ability to tailor the information environment of workers, including their engagement with virtual peers. Using a series of field experiments conducted in an online labor market, we find strong evidence that even brief exposure to peers strongly affects output. Workers exposed to high-output peers subsequently raise their own output. However, there is no "free lunch"; exposure to high-output also causes some workers to exit. Even having one worker evaluate another worker—a common quality control pattern—affects subsequent output. We show that this peer effect is not solely due to social learning about employer expectations. When we remove all uncertainty about an employer's expectations, workers still increase output, even beyond the employer's expectations when exposed to other workers doing the same.

*We thank Iwan Barankay, Raj Chetty, Lydia Chilton, Nicholas Christakis, Malcolm Wiley-Floyd, Larry Katz, Renata Lemos, and Rob Miller for helpful comments and suggestions. We thank Robin Horton, David Yerkes, and Heidi Yerkes for their help in preparing the manuscript. Horton would like to thank the NSF-IGERT Multidisciplinary Program in Inequality and Social Policy (Grant No. 0333403), the University of Notre Dame, and the John Templeton Foundation's Science of Generosity Initiative for their generous financial support. We were greatly aided by the excellent research assistance of Alex Breinen, John Comeau, Talia Goldman, Michelle Lindner, and Justin Keenan. Author contact information, datasets, and code are currently, or will be, available at http://www.john-joseph-horton.com/.

1 Introduction

The rise of online labor markets, remote work, peer production and has created many more forms of computer-mediated work. It has also given firms an increasing ability to tailor the information environment of workers, including their observability of—and interaction with—virtual peers. Even in more conventional settings, firms can add social features to existing work environments, bolstering information channels from peers (Wu, 2013). Firms can have workers work in isolation. But a common pattern for quality control in crowdsourcing is to have peers evaluate each other, thereby creating a channel for peer effects Wang et al. (2017). There is large literature about the importance of peer effects in offline production settings (Chan et al., 2014; Bandiera et al., 2010; Mas and Moretti, 2009; Falk and Ichino, 2006), and in these studies, they typically take peer effects as given. There is very little focuses on the design possibilities inherent in strong peer effects—or, in short, how to wield peer effects for the firm's ends.[1]

This paper reports the results of a series of field experiments conducted in an online labor market whose purpose was to determine how peer effects could be wielded. In each experiment, we operated as employers looking to have a series of images labeled—a common task in the marketplace (Wang et al., 2017). Each experiment required workers to decide how much costly output to produce. Some experiments asked workers to evaluate the output of their peers. This setup allowed us to both (a) measure peer-punishment and (b) expose workers to the output choices of their peers. To detect peer effects, some ex-

[1]This study stands in contrast to non-work computed mediated settings where the design powers created by controlling social information have been considered (Burtch et al., 2016; Dou et al., 2013). There have been some attempts to exploit peer effects, but it has proven challenging in practice. For example, Carrell et al. (2013) offers a cautionary tale about trying to engineer Air Force Academy peer groupings to improve academic performance. The effort strongly backfired for unforeseen reasons relating to the micro-foundations for peer effects in education. The exploitation of peer effects has had its greatest successes not in production settings, but rather on product uptake—see Aral and Walker (2014, 2011); Bapna and Umyarov (2015).

periments asked workers post-evaluation to complete another task, for which we could also measure the output.

Our first experiment established three important findings: workers were uncertain about the "firm's" expectations, they found output costly, and they responded to signals about the firm's expectations (as conveyed by employer-provided work samples). The experiment established that workers felt a need to learn about employer expectations. The second experiment introduced a channel for peer effects by having workers perform a task and then evaluate the output of a peer. The experiment showed that workers readily punished low-output/low-effort peers, even though direct free-riding externalities were not present.

The third experiment added one component to the second experiment: evaluators had to perform a follow-on labeling task. Exposure to high-output peers increased workers' output on the follow-on task. Although this experiment demonstrated that peer effects can easily be generated, it did not distinguish between social learning about employer expectations and equity-based explanations. That is, workers exposed to high output peers could have revised their beliefs about what the employer expected or revised their beliefs about what is fair (or what their peers would likely regard as fair), or both.

Our fourth experiment was designed to distinguish between social learning about employer expectations and equity-based peer effects. In that experiment, the firm's exact expectations were clearly communicated, therefore removing uncertainty about what the firm expected. This communication of standards was effective, in that on the first task, nearly all workers complied with the output level we requested. After this initial task, some evaluators evaluated work that complied with the firm's expectations, while others evaluated work that exceeded the firm's specified output expectations. Not only did evaluators not punish high-effort workers (but technically speaking, non-complying workers), but also they imitated them. They too increased output in the subsequent task beyond what the firm requested.

An interpretation of fourth experiment—and the punishment results more generally—is that workers do not simply assess output by whether it perfectly matches what the employer states it wants. It is likely that the standard applied by the evaluating workers was not the employer's standard *per se*, but rather a judgment of the effort made by each evaluated worker relative to his/her own work. Consistent with this interpretation, across experiments, punishment of peers was strongly increasing in the evaluator's own output.

Given the short duration of the "relationships" created online and the fact that "peers" never interacted in person and are completely anonymous with respect to both the employer and each other, strong peer effects seemed *a priori* unlikely. Yet they clearly emerged. What seems to matter for peer effects in productivity is the observability of output rather than the direct social interaction. Also, as each worker was working for us as an employer for the first time, we expected that social learning about our expectations would be particularly important, yet this too was not the case. In settings where interactions are longer lasting, any social learning would likely diminish in importance over time, leaving only fairness considerations.

We find that peer effects are strong and arise readily, but we also show that they could be in some cases too strong; our final experiment showed that the effects of peers was strong enough to counteract our clearly communicated output requests. Although this may seem like a free lunch—higher output at the same pay—we also observed in the first experiment that workers sorted away from our tasks when they perceived our output standards to be excessive, made it ambiguous whether actual firms would welcome the peer effects we created in the last experiment. Consistent with the view that firms would have to compensate workers for higher and thus more costly productivity, Cornelissen et al. (2013) find peer effects in *wages*, but only for relatively low-skilled occupations. One interpretation of this finding is that those low-skilled occupations have both observable output (a prerequisite for peer effects in output) and a more direct relationship between productivity and the disutility of effort.

The paper is organized as follows: Section 2 describes the empirical setting for our experiments. Section 3 presents the experimental designs and describes the experimental tasks and the features common across experiments. Section 4 presents the results, and Section 5 concludes.

2 Empirical setting

The experiments were conducted on Amazon's Mechanical Turk (MTurk), an online labor market. Online labor markets have themselves attracted considerable research attention in IS(Hong et al., 2015; Hong and Pavlou, 2017; Kanat et al., 2018).[2] 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.[3] Some examples in economics include Mason and Watts (2009), Barankay (2010), Chandler and Kapelner (2013), and Horton and Chilton (2010). Online experiments offer significant advantages relative to most laboratory experiments in cost, speed of accruing subjects, and representativeness of the subject pool. However, they tend to be harder to control than conventional laboratory experiments. Despite this difficulty, Horton et al. (2011) demonstrate that valid causal inferences can be drawn in online labor markets such as MTurk.

Tasks posted on MTurk are called Human Intelligence Tasks (HITs). HITs vary, but most are small, simple tasks that are difficult or impossible for machines but easy for people, such as transcribing audio clips, classifying and tagging images, reviewing documents, and checking websites for pornographic content. The originators of HITs are called "requesters." Requesters and workers are anonymous to each other. Rarely do they interact on a repeat basis. The requester constructs the user interface for the HIT and sets the conditions for

[2]Interestingly, the platforms themselves are heavy users of peer-evaluation—a channel for peer effects, which we show–as post transaction rating systems are key to solving informational asymmetries (Moreno and Terwiesch, 2014).

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

payment, worker qualifications, and timing (for example, how long a worker can work on a task).

To become eligible, a would-be worker must have a bank account and must have created a profile with Amazon.[4] Workers can only have one account, and Amazon uses several technical and legal means to enforce this restriction. Workers can readily see 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. Once they accept a HIT, they can begin work immediately.

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. This ability of requesters to reject work deters work that falls short of employer expectations. Requesters may also pay bonuses, allowing tailored payments based on individual performance within a nominally piece-rate HIT.

3 Experimental design

Our first experiment, EXP-BASELINE, addressed two questions: 1) Do workers view the task as costly?, and 2) Does providing output samples effectively convey employer expectations, as revealed by changes in output following observation of those samples? The experiment EXP-PUNISH addressed a third question: 3) Will workers punish low effort/low productivity outcomes relative to some standard? This test was conducted by exposing workers to either a high- or low-output peer, and then asking the evaluating worker (a) whether we, as the employer, should "approve," that is, accept and pay for the work, and (b) how the evaluator wanted to split a bonus with the evaluated worker. The experiment EXP-PEERS, addressed the question: 4) Does a worker's exposure to the output choice of a peer—shown via an evaluation—influence that worker's

[4]MTurk workers are often called "Turkers." Their ranks appear to be split approximately evenly between the US and India. Horton (2011) finds workers generally view online employers as having the same level of trustworthiness as offline, conventional employers. For the demographics of the MTurk population, see Ipeirotis (2010).

subsequent output?

In the experiment EXP-EXPLICIT, we attempted to remove any uncertainty about the firm's expectations in order to test: 5) Will workers thus informed punish work that exceeds the employer's stated expectations, and therefore does not conform to it? In other words, will evaluators punish workers for not conforming to what was expected? A worker might be understandably reluctant to punish high-effort work even if it failed to comply on a technicality, especially if they think the employer will be pleased with or has free disposal on "excess" output. The worker might believe as well that they are making the same decision the employer would have made in the same context. However, the interesting result of this experiment is whether exposure to this high-effort but non-complying work affects the evaluator's subsequent output. If it does, then it suggests a channel by which peer pressure can sustain behaviors or levels of output that deviate from what the firm claims it wants.

3.1 Preliminaries and common elements across experiments

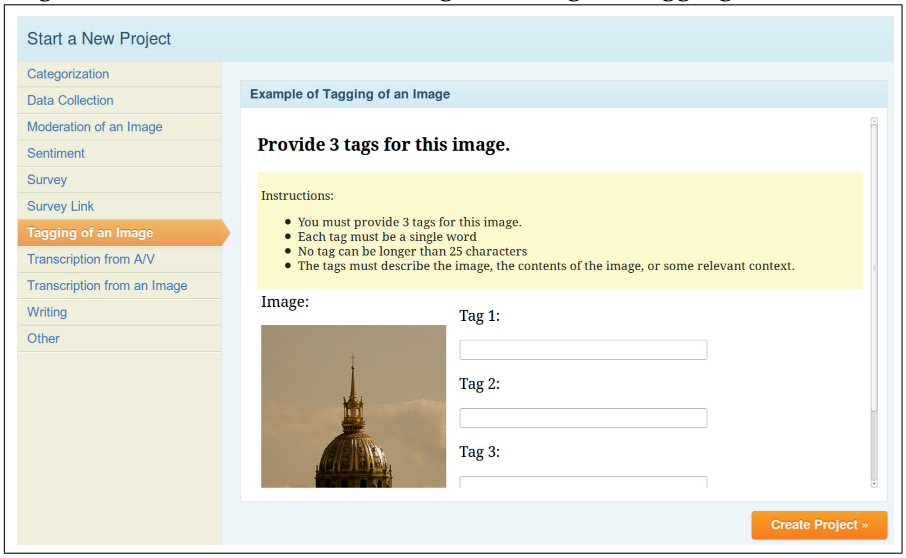
In every experiment, before agreeing to participate, would-be workers were told about the task and the payment for it, and were shown a completed work sample. The work sample was a "screenshot" of the image-labeling interface as it would look after a worker completed the task.[5]

3.1.1 Task and interface

Computers have a difficult time recognizing objects in images, yet this task is often valuable for firms. Thus, image labeling is a common "human computation" task found on MTurk (von Ahn and Dabbish, 2004; Huang et al., 2010). It is one of a handful of canonical MTurk tasks for which Amazon has created pre-made templates. Figure 1 depicts one of Amazon's pre-made interfaces.

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

Figure 1: Pre-made Amazon image labeling or "tagging" interface



Notes: Screenshot of the image-tagging interface developed by Amazon.

To label images, workers in our experiment used an interface we developed, shown in Figure 2. To add a label, workers clicked a button labeled "Add a label" positioned below the image.[6] Clicking the button brought forth a new blank for the workers to fill in. Workers could add as many labels as they wanted. When they were finished, they clicked a button labeled "Submit labels."

3.1.2 Peer evaluations

A worker played two roles: the producer of labels of images and the evaluator of the work of a peer who had also engaged in the image-labeling task (except in EXP-BASELINEwhich had no evaluation step). To evaluate the peer, the evaluating worker viewed the labeling worker's assigned image and that worker's generated labels. The evaluating worker then made recommendations on two matters: whether to approve (pay for) or not approve (not pay for) the task, and

[6]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 (for example, photos depicting elaborate meals with many easily recognizable different foods).

how to split a 9-cent bonus between him- or herself and the evaluated worker.

All evaluating workers within an experimental group assessed the same worker's output. This evaluated worker was chosen at random from subjects who had participated in previous experiments and whose work had exhibited the desired property for that experimental group (either high or low effort). The workers performing either task—labeling images or evaluating performance—-likely regarded the evaluation and bonus schemes as unremarkable, since a very common quality-control technique is to have workers evaluate the work of other Turkers. Furthermore, bonuses are frequently used to incentivize good performance.

For the accept/reject question, the evaluating workers were asked:

> Should we approve this work?

They had to answer "yes" or "no." For the peer evaluation, the evaluators 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.

The evaluating workers selected an answer from a list of 9 options of the form "X cents for the other, $9 - X$ 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). Both questions were asked on the same survey page, and the evaluators could answer them in either order, though the approval question appeared first on the page. At the end of the experiment, we implemented all choices, with the recommended bonuses paid to the evaluated workers and their evaluators, in accordance with the evaluator's preferences.

3.1.3 Demographic survey and allocation

In each experiment, subjects answered a short demographic survey before beginning work. Subjects were asked their sex, nationality, and whether they were doing this work primarily to earn money, learn new skills, or have fun. Demographics differed slightly across experiments, probably due to differences in the times the experiments were launched. Although one might conjecture that the survey would raise suspicions that the task was an experiment, we view this as unlikely. Asking workers for basic demographic information is fairly common in the market.[7] In each experiment, subjects were assigned alternately to groups in order of arrival time (for instance, Subject 1 was assigned to treatment, Subject 2 to control, Subject 3 to treatment, and so on) to give better balance.

3.2 EXP-BASELINE: Baseline experiment

Participants in this experiment were assigned to either HIGH or LOW. In HIGH, the employer-provided work sample showed 9 labels, compared to only 2 labels in LOW. Figure 2 shows the two work samples. After viewing their particular work sample, workers chose to participate and label an image or to exit. Those who chose to participate then performed a labeling task.

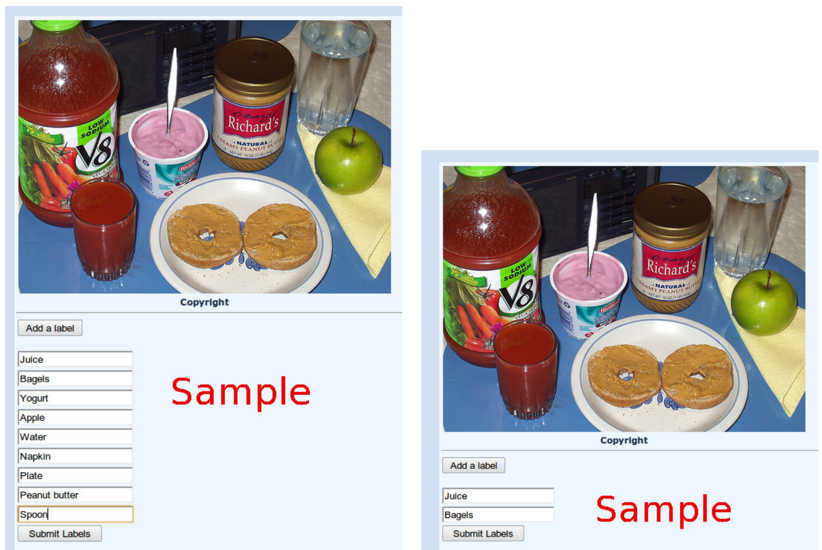
Table 1 reports the means of various demographic measures collected for the EXP-BASELINE participants. While the set of covariates is limited, there is no indication that the randomization was ineffective. The same analysis was conducted for the other experiments reported in the paper; the results mirror those from the first experiment. Although not reported, the full dataset and this auxiliary analysis is available online.

3.3 EXP-PUNISH: Punishment experiment

The job posting for this experiment was the same as in EXP-BASELINE with one addition: potential subjects were told that they would be evaluating the work of another worker. Before accepting the task, all subjects were shown the HIGH

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

Figure 2: Work samples shown to workers prior to task acceptance in EXP-BASELINE



(a) HIGH work sample (b) LOW work sample

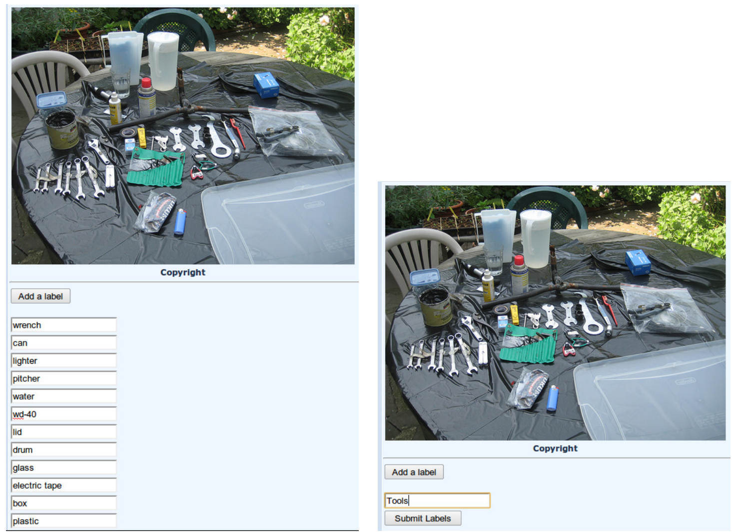
Notes: The panels show the work sample given to would-be workers considering accepting the image-labeling task. Subjects assigned to HIGH were shown the left image with its 9 labels; subjects assigned to LOW were shown the right image with its 2 labels.

Table 1: Comparison of covariate pre-treatment means for exerimental groups in EXP-BASELINE

Variable	HIGH group	LOW group	t-stat
From India	0.54	0.43	1.13
Male	0.61	0.55	0.54
Reports that primary motivation is money	0.78	0.74	0.43
From the US	0.35	0.40	-0.56

Notes: This table reports the means for the HIGH and LOW experimental groups in EXP-BASELINE. The t-statistic for a difference in means is reported.

Figure 3: Work evaluated by subjects in EXP-PUNISH



(a) GOOD work sample (b) BAD work sample

Notes: The image in the left panel is the work output evaluated by workers assigned to GOOD. The image in the right panel is the work output evaluated by workers assigned to BAD.

work sample displayed in Figure 2. To reward the additional evaluation work, the participation payment was raised from 30 cents to 40 cents. The requested sample size was also increased from 100 to 200.

In this experiment, all subjects first completed the same image-labeling task before they were randomized into the two groups. Subjects assigned to GOOD inspected the output of a worker (selected from EXP-BASELINE) who had produced 12 unique labels. Subjects assigned to BAD inspected the output of a worker who had produced only one unique label. The output samples of the evaluated workers are shown in Figure 3.

3.4 EXP-PEERS: Peer experiment

EXP-PEERS was identical to EXP-PUNISH with one addition: we asked workers to do a second labeling task after completing their evaluation task. The two experimental groups are GOOD and BAD. In GOOD, subjects evaluated a worker who produced 11 labels; in BAD, subjects evaluated a worker who produced only 2 labels. The requested sample size was 300, and each subject's payment was 40 cents.

3.5 EXP-EXPLICIT: Explicit experiment

EXP-EXPLICIT re-ran EXP-PEERS with one exception: we provided explicit employer instructions to subjects before exposing those workers to the output of peers. The workers were told that they should produce 2 labels per image. The requirement of 2 labels was stated before workers began the task, and was presented again with each of the two image-labeling tasks, directly above the image and in the instructions. The left panel of Figure 4 shows initial instructions, while the right panel shows the instructions placed above both the first and the subsequent labeling tasks. In the left panel, instructions explain that the worker is to provide two labels for an image, rate the work of another worker and then provide two labels for an additional image. The right panel shows the interface for the subsequent labeling task. Note that it reiterates the requirement that the worker provide two labels.

After performing the initial task, workers were randomized into one of two groups: MEET, in which subjects evaluated a work sample showing $x = 2$, and EXCEED, in which subjects evaluated a work sample showing $x = 11$. After evaluating the work, subjects performed an additional image-labeling task. The requested sample size was 300, and the payment was 40 cents.

Table 2: Overview of the experiments

Description	Results
EXP-BASELINE (Groups = HIGH & LOW, Payment = 30¢, N = 93)	
Subjects, according to experimental group, viewed one of two employer-provided work samples, then produced whatever number of labels they chose (if any). Work samples differed by experimental group. In HIGH, subjects first viewed a high-output work sample (many labels). In LOW, subjects first viewed a low-output work sample (few labels).	HIGH increased labor supply on intensive margin, but decreased it on the extensive margin.
EXP-PUNISH (Groups = GOOD & BAD, Payment = 40¢, N = 167)	
Subjects viewed an employer-provided high-output work sample, then produced whatever number labeles they chose. Subjects then evaluated another worker's work product. Subjects in GOOD evaluated a high-effort work sample. Subjects in BAD evaluated a low-effort work sample.	GOOD subjects decreased punishment. More productive workers punished more.
EXP-PEERS (Groups = GOOD & BAD, Payment = 40¢, N = 275)	
Subjects viewed an employer-provided work sample, then chose how many labels to produce. Subjects then evaluated another worker's work product, then labeled a second image. Subjects in GOOD evaluated a high-effort work sample. Subjects in BAD evaluated a low-effort work sample.	GOOD raised output on the second task. Effects were stronger for subjects with high effort in the first task.
EXP-EXPLICIT (Groups = MEET & EXCEED, Payment = 40¢, N = 272)	
Subjects viewed an employer-provided work sample with 2 labels. Subjects were told that 2 and only 2 labels should be produced. Subjects then labeled an image, evaluated another worker's work product, then labeled a second image, with the same instruction to create only 2 labels. In EXCEED, subjects evaluated a worker producing too many labels. In MEET, subjects evaluated a worker producing the required number of labels.	EXCEED caused many workers to not comply, even when they complied on the first task. Workers did not punish high-effort but non-complying work.

Figure 4: Exp-Explicit communication of employer standards

(a) Initial Exp-Explicit instructions

(b) Exp-Explicit instructions repeated

Notes: This screenshot shows the labeling interface in Exp-Explicit. The left panel shows the initial instructions which explain that the worker is to provide 2 labels for an image, rate the work of another worker and then provide 2 labels for an additional image. The right panel shows the interface for the subsequent labeling task. Note that it reiterates the requirement that the worker provide 2 labels.

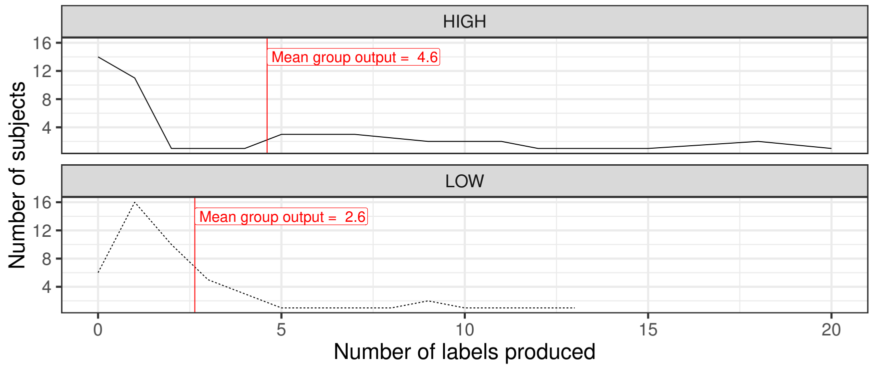
4 Results

4.1 Exp-Baseline: Is the labeling task perceived by the workers as costly, and are the conveyed employer expectations salient?

The first experiment shows that having a worker observe employer-provided work samples affects labor supply on the extensive margin. Showing a high-work sample in High without requiring any peer evaluations effectively conveys the firm's expectations.

The main results from Exp-Baseline are displayed in Figure 5. In the figure, the number of labels produced is on the x-axis, while the number of subjects producing that many labels is on the y-axis. The top facet shows the output for subjects in the High group, while the bottom facet shows output for subjects

Figure 5: Distribution of labels produced by experimental group in EXP-BASELINE



Notes: This figure plots the count of experimental subjects producing the number of labels listed on the x-axis. For example, 14 subjects in the HIGH experimental group produced 0 labels. Subjects in HIGH 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. Group mean output is shown, with zeros included.

in LOW. It shows that 5 subjects in HIGH produced more than 12 labels, but only 1 subject in LOW produced more than 12 labels.

It also reveals that many more subjects in HIGH produced no output at all, namely 14 verus 5. Presumably, they were discouraged by the expectations implied in the image they were shown. In Table 3, we confirm what Figure 5 suggests: In Column (1), we regress whether the worker produced any labels at all on the treatment indicator, while in Column (2), we regress the count of labels on that same indicator.

HIGH employer expectations reduced labor supply on the extensive margin, but increased it on the intensive margin. However, the latter effect outweighed the former: even with the non-participants included as $x = 0$ observations, subjects in HIGH produced roughly 2 more labels per person, on average (4.6 versus 2.6). Because this unconditional output rose significantly in HIGH, we know that selection does not explain the observed increase in output.

This experiment has important implications for our study. These findings indicate that we chose a task: (a) that workers found personally costly to per-

Table 3: Effects of perceptions of employer's output expectations on extensive and intensive labor supply in EXP-BASELINE

	Dependent variable:	
	Any output?	Output
	(1)	(2)
Assigned to HIGH	-0.177^{*}	1.970^{*}
	(0.084)	(0.939)
Intercept	0.872^{***}	2.638^{***}
	(0.059)	(0.660)
Observations	93	93
R^2	0.046	0.046
Adjusted R^2	0.036	0.036

Notes: This table reports the results of two robust OLS regressions where the dependent variables are measures of worker labor supply. In this experiment, all subjects were invited to participate in a paid image-labeling task. Those subjects assigned to HIGH viewed an employer-provided work sample with 9 labels, while subjects in LOW viewed a sample with only 2 labels. In Column (1), the outcome variable is labor supply on the extensive margin, that is, whether the worker accepted the task and generated any labels at all. In Column (2), the outcome variable is labor supply on the intensive margin, that is, the number of labels the worker produced, with 0s included. Significance indicators: $p \leq 0.05 : *$, $p \leq 0.01 : **$ and $p \leq .001 : ***$.

form, in a setting where (b) the employer's expectations were easily conveyable, and (c) that workers factored those expectations into their decision making.

4.2 EXP-PUNISH: Do workers punish low effort/productivity?

In EXP-PUNISH, workers were randomly assigned to evaluate either good work (in GOOD) or bad work (in BAD). We wanted to test whether evaluators would still punish even in the absence of direct free-riding externalities.

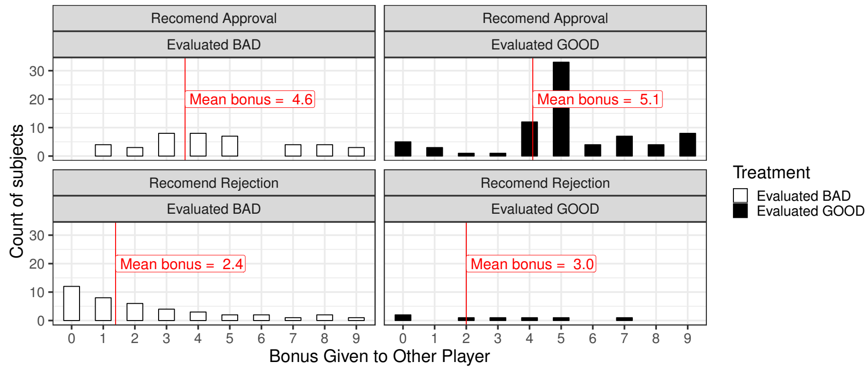
The main results from EXP-PUNISH can be seen in Figure 6. The figure contains four histograms; each shows the count of evaluators choosing different allocations of the 9-cent bonus between themselves and their evaluated worker. The plots are defined by experimental group (column) and evaluator recommendation regarding approval (row). Evaluators in GOOD were very unlikely to recommend rejection, whereas evaluators assigned to BAD were fairly likely to do so. For the BAD/reject evaluators, the modal transfer was 0 cents. Apart from the evaluators in BAD who recommended rejection, few evaluators in either group transferred 0 cents.

Most GOOD evaluators, as well as a large number of evaluators who recommended approval despite being in BAD, chose a more or less even split that gave 4 or 5 cents to the worker.

Table 4 confirms the results the graphical analysis portrays. The dependent variable in Columns (1) and (2) indicates whether the evaluating worker recommended that we pay the evaluated worker. The independent variable in Column (1) is the treatment indicator. Evaluators assigned to GOOD were far more likely to recommend acceptance than those assigned to BAD. The baseline was 50% in BAD and increased to above 90% in GOOD.

The independent variable in Column (2) is the number of labels the evaluating workers produced themselves, prior to the evaluation. As is evident, workers who produced more labels themselves were more likely to recommend rejection. The dependent variable in Columns (3) and (4) is the bonus size. As with the accept/reject measure, we can see that GOOD increased transfers in

Figure 6: Distribution of reward and punishment by workers of their peers based upon perceived effort/quality in EXP-PUNISH



Notes: This figure shows the bonus split and the accept/reject recommendation by treatment group. Subjects in BAD evaluated work with 1 generic label, while subjects in GOOD evaluated work with 12 specific and appropriate labels. Group mean bonus size is shown with a vertical line.

the contextualized dictator game. Moreover, these transfers were decreasing in the evaluator's own output. For both measures of punishment, higher-output workers were harsher critics; they were both more likely to recommend rejection and to award smaller bonuses.

4.3 EXP-PEERS: Do peers influence output?

As in EXP-PUNISH, subjects in GOOD in EXP-PEERS were more likely to recommend acceptance and transfer larger bonuses on average than subjects in BAD. This was expected, and we do not present the analysis. Rather, we focus on the evaluating worker's subsequent output. Recall that, in the EXP-PEERS experiment, after the initial output task and evaluation task, workers performed a second labeling task. As we show, workers assigned to GOOD produced more output than those in BAD in their second production task.

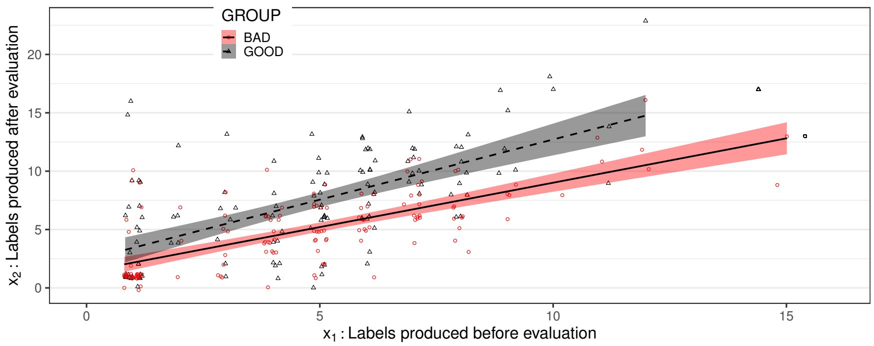
Figure 7 plots the workers' subsequent output, x_2, versus their initial output, x_1. We fit separate lines for the two treatment groups, with 95% confidence intervals for the conditional mean illustrated with shaded regions. The regres-

Table 4: Reward and punishment by workers of their peers based upon perceived effort/output in EXP-PUNISH

	Dependent variable:			
	Recommend Approval?		Bonus Size Awarded	
	(1)	(2)	(3)	(4)
Assigned to GOOD	0.418^{***}		1.442^{***}	
	(0.063)		(0.390)	
Initial Output, x_1		-0.055^{***}		-0.259^{***}
		(0.013)		(0.074)
Intercept	0.500^{***}	0.960^{***}	3.488^{***}	5.376^{***}
	(0.045)	(0.065)	(0.278)	(0.383)
Observations	167	167	167	167
R^2	0.213	0.105	0.077	0.069
Adjusted R^2	0.208	0.100	0.071	0.064

Notes: This table reports the results of robust OLS regressions of two measures of peer punishment in EXP-PUNISH. In GOOD, workers evaluated the work output of a worker generating 12 specific and highly appropriate labels; the other workers in BAD evaluated a worker producing only 1 generic label. In Columns (1) and (2), the outcome measure is whether the worker recommended that the employer "approve" the work of the evaluated worker and thus pay them. In Columns (3) and (4), the outcome measure is the amount of bonus transferred to that evaluated worker, out of an endowment of 9 cents. The key independent variable in Columns (1) and (3) is the treatment indicator, GOOD, whereas in Columns (2) and (4), the key independent variable is the *evaluating* worker's output. The regressions in (2) and (4) are not causal, but they illustrate the strong negative relationship between own-output and the tendency to punish. Significance indicators: $p \leq 0.05 : *$, $p \leq 0.01 : **$ and $p \leq .001 : ***$.

Figure 7: Subsequent number of labels, x_2, versus initial number of labels, x_1, by whether subject evaluated GOOD or BAD in EXP-PEERS



Notes: In this plot the y-axis is subsequent output, x_2, and the x-axis is initial output, x_1, in EXP-PEERS, by experimental group. All subjects performed an identical initial task and chose some number of labels to provide (shown on the x-axis). Each subject then evaluated another subject's work that demonstrated either (BAD) low productivity or (GOOD) high productivity. All output levels are randomly perturbed by a small amount to prevent over-plotting.

sion line for GOOD lies everywhere above the line for BAD, and it is steeper. The regression analysis in Table 5 confirms this graphical analysis. The outcome measure in each column is the number of labels each evaluating worker produced in the second labeling task. Column (1) shows that assignment to GOOD increased the mean number of labels by more than 2, from a baseline of just 5. Column (2) adds the worker's pre-treatment output as a regressor. As would be expected, past performance predicts future performance. Column (3) adds an interaction term between the number of number of labels, x_1, and the treatment indicator. The positive and significant effect of the interaction term implies that the initially high-output workers had the greatest subsequent output response to their exposure.

Table 5: Effects of evaluating a co-worker on subsequent output in EXP-PEERS

	Dependent variable:		
	Number of labels produced after evaluation, x_2		
	(1)	(2)	(3)
Assigned to GOOD	2.169^{***}	2.290^{***}	0.971
	(0.488)	(0.370)	(0.699)
Initial Output, x_1		0.884^{***}	0.764^{***}
		(0.063)	(0.083)
GOOD $\times x_1$			0.280^{*}
			(0.126)
Intercept	5.043^{***}	0.815^{*}	1.388^{**}
	(0.338)	(0.395)	(0.470)
Observations	265	265	265
R^2	0.070	0.469	0.479
Adjusted R^2	0.066	0.465	0.473

Notes: This table reports robust OLS regressions from EXP-PEERS. Workers were randomly assigned to evaluate either a worker exhibiting (GOOD) high productivity/effort or (BAD) low productivity/effort. In each column, the output measure is the number of labels a worker produced after evaluating another worker, x_2. Significance indicators: $p \leq 0.05 : *$, $p \leq 0.01 : **$ and $p \leq .001 : ***$.

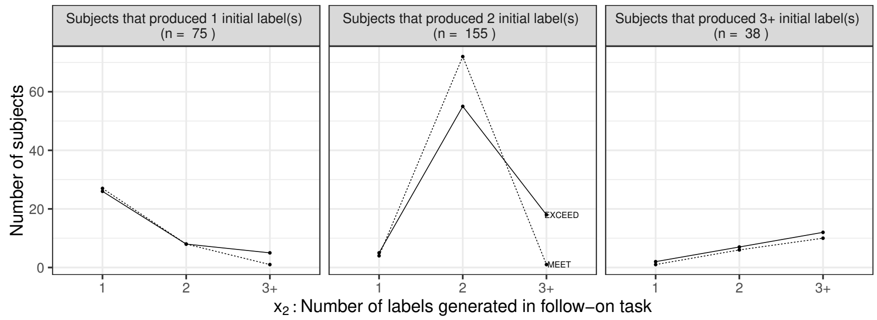
4.4 EXP-EXPLICIT: How do peer effects change when employer requirements are explicitly identified? Do evaluators punish high effort if it deviates from employer expectations?

The results from EXP-PEERS show that workers are influenced by peers and not simply by employer-provided sources of information. However, these results do not establish the source of these peer effects. Are they driven by social learning about the firm's standards or by social learning about peer standards?[8] EXP-EXPLICIT was designed to distinguish between these two potential sources of social learning. In EXP-EXPLICIT, workers first did a task in which they were given explicit instructions to produce only two labels. They then evaluated a peer who either precisely met the employer's expectation (MEET) or who produced more than two labels (EXCEED). After this evaluation, the evaluating worker performed another labeling task. For this subsequent labeling task, our key question was whether having explicit employer instructions "stamped out" peer effects. Note that this experiment offers a strong test, since the employer imposes a ceiling, as well as a floor, on production. The explicit employer instructions are designed to let the worker know the firm's precise expectation, potentially sealing the social learning channel for peer effects.

Figure 8 illustrates the main results of this experiment but needs some explanation. The plots show subjects moving from different pre-treatment output "bins" to post-treatment output bins. The three side-by-side panels correspond to the three output levels that workers chose for the initial task: $x_1 = 1$, $x_1 = 2$, and $x_1 \geq 3$ (labeled 3+ in the figure). Thus, the left panel consists of only those subjects that generated $x_1 = 1$ label; the center panel consists of only those subjects that initially complied by producing $x_1 = 2$ labels; and so on. Within each panel, these same bands are used again for the x-axis, for the count of labels produced in the subsequent task; that is, $x_2 = 1$, $x_2 = 2$, and $x_2 \geq 3$. The y-axis indicates the count of subjects in that (x_1, x_2) bin. The two experimental groups

[8] Note that in defining what the firm expects for output, we use the terms "expectation," "requirement" and "standard" interchangeably.

Figure 8: Numbers of workers in different output bands for the subsequent task in EXP-EXPLICIT, faceted by initial output, with experimental groups indicated by line type



Notes: For initially complying subjects (middle panel), assignment to EXCEED increased the relative number of subjects producing additional (and hence non-complying) output in the second task.

are shown separately by differences in line type: MEET is shown with a dashed line; EXCEED is shown with a solid line.

Several interesting results emerge in Figure 8. First, the center panel received most of the observations, indicating that the majority of workers initially complied with employer requirements and produced exactly 2 labels. In the left and right panels, there are no differences between the experimental groups; the lines nearly overlap. The center panel shows results that are quite different. For these subjects who complied initially, exposure to EXCEED reduced compliance in the subsequent task. Specifically, the EXCEED treatment increased output among those initially complying. (The dotted line lies below the solid line at $x_2 \geq 3$.)

Table 6 confirms our graphical analysis. However, unlike in Figure 8, it restricts attention solely to those subjects who originally complied, $x_1 = 2$. In this restricted sample, the chance that worker error or misunderstanding drove the results is reduced. Subjects in MEET had a 90% compliance rate, whereas subjects in EXCEED had only a 70% compliance rate. Note that we can restrict our sample in this way because the x_1 choice was made pre-treatment. In Columns

(1) and (2), the outcome variable is an indicator for compliance with the employer instructions, $\mathbb{1}\{x_2 = 2\}$, in the subsequent task. In Column (3), the outcome variable is an indicator for choosing $\mathbb{1}\{x_2 \geq 3\}$.

We can see in Column (1) that assignment to MEET increased compliance in the subsequent task. In Column (3), we can see where the noncompliers went; they almost universally increased their output to $x_2 \geq 3$. Might these results be driven by a misunderstanding of the instructions? The regression in Column (2) helps us answer that question. In Column (2), we interact the treatment indicator with an indicator for whether the subject was from India, where English is often not the primary language. Although it is not necessarily the case that subjects from India have worse English skills than those from the US, in other online labor markets, employers at least act is if they expect workers from less developed countries to be more likely to have communication difficulties (Agrawal et al., 2013). In our regressions, we see no country-specific effect, which suggests that exposure to the high-output peer did not cause the subject to simply second-guess the employer's written instructions.

5 Conclusion

One contribution of our findings is to offer yet another example of peer effects in a real production setting. However, our evidence goes beyond mere existence, showing that these effects are strong and arise in a setting when there is no personal interaction between peers. Across the experiments, variation in the exposure to different work samples explained a substantial fraction of the variation in observed output. Moreover, this finding arose despite an experimental, online setting that offered very short "interactions" that were ostensibly one-shot.

A natural question for managers is whether they should encourage peers to influence each other, such as by optimally arranging teams to maximize productivity. Context surely matters greatly. In settings where effort and productiv-

Table 6: Effects of exposure to high-effort, non-complying peer work after explicit employer instructions about compliance

	Dependent variable:		
	Comply		Over-Produce
	$\mathbb{1}\{x_2 = 2\}$	$\mathbb{1}\{x_2 = 2\}$	$\mathbb{1}\{x_2 > 2\}$
	(1)	(2)	(3)
Assigned to MEET	0.230^{***}	0.247^{**}	-0.218^{***}
	(0.059)	(0.077)	(0.050)
India		0.046	
		(0.086)	
Complied ($x_1 = 2$) × India		−0.045	
		(0.122)	
Constant	0.705^{***}	0.688^{***}	0.231^{***}
	(0.042)	(0.054)	(0.035)
Observations	155	155	155
R^2	0.089	0.091	0.110
Adjusted R^2	0.083	0.073	0.104

Notes: This table reports robust OLS regressions from EXP-EXPLICIT. In this experiment, subjects were assigned to evaluate either (MEET) a work sample exactly meeting the employer's expectations or (EXCEED) a non-complying (but high-effort) sample. In Columns (1) and (2), the outcome variable indicates whether the worker complied with the employer's stated output requirement of exactly two labels, $\mathbb{1}\{x_2 = 2\}$. In Column (3), the outcome is whether the worker exceeded the employer's stated output requirement of 2 labels. Significance indicators: $p \leq 0.05 : *$, $p \leq 0.01 : **$ and $p \leq .001 : ***$.

ity are tightly coupled and workers can easily monitor each other, peer pressure would seem to provide a kind of free lunch for the firm. However, recall from our first experiment that workers who infer that the firm had high standards were more likely to exit and complete no labels, so perhaps “cheap lunch” is a more appropriate characterization, in that our per-output costs went down substantially because of the productivity boost from peer effects.

In some contexts, giving workers the ability to punish or reward their peers may hurt, not help. Workers might enforce inefficient standards. For example, in noise-filled environments, where the connection between effort and productivity is tenuous and where explicit firm-provided incentives are highly muted, the firm might be worried that otherwise-good workers would feel compelled to feign industry to placate effort-monitoring peers who punish and/or reward. When such dangers loom, the firm might even want to go so far as to prevent workers from monitoring each other. In short, enabling peer pressures when peers may be punishing or rewarding undesired outcomes is risky.

One unexplored organizational implication of these findings is the possibility of a feedback loop or cascade. The potentially causal dependence between one’s own productivity and the willingness to punish, combined with susceptibility to peer effects, provides one such mechanism. Cascades can be harmful, for example, if after workers observe idiosyncratically bad work, they lower their own output and punish less, which in turn reduces other workers’ incentives to be highly productive. On the beneficial side, employers will seek to harness peer effects when there is strong potential for a constructive cascade, when idiosyncratically productive work spurs superior output from those monitoring, who then induce superior output from others. The potential for hard effort to spread from one worker to another helps to 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.

References

Agrawal, Ajay K, Nicola Lacetera, and Elizabeth Lyons, "Does information help or hinder job applicants from less developed countries in online markets?," Technical Report, National Bureau of Economic Research 2013.

Aral, Sinan and Dylan Walker, "Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks," *Management Science*, 2011, *57* (9), 1623–1639.

_ **and** _ , "Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment," *Management Science*, 2014, *60* (6), 1352–1370.

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

Bapna, Ravi and Akhmed Umyarov, "Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks," *Management Science*, 2015, *61* (8), 1902–1920.

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

Burtch, Gordon, Anindya Ghose, and Sunil Wattal, "Secret admirers: An empirical examination of information hiding and contribution dynamics in online crowdfunding," *Information Systems Research*, 2016, *27* (3), 478–496.

Carrell, Scott E, Bruce I Sacerdote, and James E West, "From natural variation to optimal policy? The importance of endogenous peer group formation," *Econometrica*, 2013, *81* (3), 855–882.

Chan, Tat Y., Jia Li, and Lamar Pierce, "Compensation and Peer Effects in Competing Sales Teams," *Management Science*, 2014, *60* (8), 1965–1984.

Chandler, Dana and Adam Kapelner, "Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets," *Journal of Economic Behavior & Organization*, 2013, *90*, 123–133.

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg, "Peer Effects in the Workplace," 2013.

Dou, Yifan, Marius F Niculescu, and DJ Wu, "Engineering optimal network effects via social media features and seeding in markets for digital goods and services," *Information Systems Research*, 2013, *24* (1), 164–185.

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

Frei, Brent, "Paid Crowdsourcing: Current State & Progress toward Mainstream Business Use," *Produced by Smartsheet.com*, 2009.

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

Hong, Yili and Paul A Pavlou, "On buyer selection of service providers in online outsourcing platforms for IT services," *Information Systems Research*, 2017, *28* (3), 547–562.

_ **, Chong Wang, and Paul A Pavlou**, "Comparing open and sealed bid auctions: Evidence from online labor markets," *Information Systems Research*, 2015, *27* (1), 49–69.

Horton, John, "Online Labor Markets," *Internet and Network Economics*, 2010, pp. 515–522.

Horton, John J., "The condition of the Turking class: Are online employers fair and honest?," *Economics Letters*, 2011, *111* (1), 10–12.

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

—, **David G. Rand, and Richard J. Zeckhauser**, "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics*, 2011, *14* (3), 399–425.

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, Panagiotis G., "Demographics of Mechanical Turk," *Working Paper*, 2010.

Kanat, Irfan, Yili Hong, and TS Raghu, "Surviving in Global Online Labor Markets for IT Services: A Geo-Economic Analysis," *Information Systems Research*, 2018.

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

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

Moreno, Antonio and Christian Terwiesch, "Doing business with strangers: Reputation in online service marketplaces," *Information Systems Research*, 2014, *25* (4), 865–886.

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.

Wang, Jing, Panagiotis G Ipeirotis, and Foster Provost, "Cost-effective quality assurance in crowd labeling," *Information Systems Research*, 2017, *28* (1), 137–158.

Wu, Lynn, "Social network effects on productivity and job security: Evidence from the adoption of a social networking tool," *Information systems research*, 2013, *24* (1), 30–51.