

Essays in Online Labor Markets

by

Dana Chandler

A.B. Economics, University of Chicago (2006)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

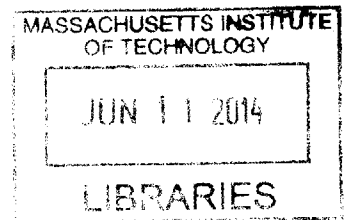
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

©2014 Dana Chandler. All rights reserved.

ARCHIVES



Signature redacted

Author

Department of Economics

Signature redacted

May 15, 2014

Certified by

David Autor

Signature redacted

Professor of Economics

Thesis Supervisor

Certified by

Heidi Williams

Assistant Professor

Thesis Supervisor

Signature redacted

Accepted by

Michael Greenstone

3M Professor of Environmental Economics

Chairman, Departmental Committee on Graduate Studies

Essays in Online Labor Markets

by

Dana Chandler

Submitted to the Department of Economics
on May 15, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis explores the economics of online labor markets. The first paper evaluates a market intervention that sought to improve efficiency within the world's largest online labor market. The second paper provides an illustration of how online labor markets can serve as a platform for helping researchers study economic questions using natural field experiments. The third paper examines the role of supervision within a firm using detailed productivity data.

In the first paper, we report the results of an experiment that increased job application costs in an online labor market. More specifically, we made it costlier to apply to jobs by adding required questions to job applications that were designed to elicit high-bandwidth information about workers. Our experimental design allows us to separate the effect of a costly ordeal vs. the role of information by randomizing whether employers see workers' answers. We find that our ordeal reduced the number of applicants by as much as 29% and reduced hires by as much as 3.6%. Overall, the applicant pool that underwent the ordeal had higher earnings and hourly wages, but not better past job performance. The ordeal also discouraged non-North American workers. We find no evidence that employers spent more when vacancies were filled, but some evidence that employer satisfaction improved. These improvements were the result of information provision rather than selection. Finally, we did not find any heterogeneity in outcomes across job category, contract types, or employer experience.

In the second paper, we conduct the first natural field experiment to explore the relationship between the "meaningfulness" of a task and worker effort. We employed over 2,500 workers from Amazon's Mechanical Turk (MTurk), an online labor market, to label medical images. Although given an identical task, we experimentally manipulated how the task was framed. Subjects in the *meaningful* treatment were told that they were labeling tumor cells in order to assist medical researchers, subjects in the *zero-context* condition (the control group) were not told the purpose of the task, and, in stark contrast, subjects in the *shredded* treatment were not given context and were additionally told that their work would be discarded. We found that when a task was framed more meaningfully, workers were more likely to participate. We also found that the meaningful treatment increased the quantity of output (with an insignificant change in quality) while the shredded treatment decreased the quality of output (with no change in quantity). We believe these results will generalize to other short-term labor markets. Our study also discusses MTurk as an exciting platform for running natural field experiments in economics.

In the third paper, we investigate whether greater supervision translates into higher quality work. We analyze data from a firm that supplies answers for one of the most popular question-and-answer (“Q&A”) websites in the world. As a result of the firm’s staffing process, the assignment of supervisors to workers is as good as random, and workers are exposed to supervisors who put forth varying degrees of “effort” (a measure based on a supervisor’s propensity to correct work). Using this exogenous variation, we estimate the net effect of greater supervision and find that a one-standard-deviation increase in supervisor effort reduces the number of bad answers by between four and six percent. By decomposing the total effect into the separate effects on corrected and uncorrected answers, we conclude that supervisor effort tends to *lower* the number of good answers among uncorrected answers. Interestingly, observable worker behaviors (i.e., answer length and time to answer a question) seemed unaffected by supervision. None of the results vary with worker experience.

Thesis Supervisor: David Autor
Title: Professor of Economics

Thesis Supervisor: Heidi Williams
Title: Assistant Professor

To my parents, for raising me in strict accordance with the principles of laissez-faire parenting. While any parent can raise their children in their own image, you accomplished something much greater. You gave me the character andchutzpah to envision my own ideal and follow my own path. My individuality and perception of the world is due entirely to you.

To the late Sheridan Wilson-Grenon, who I will always miss, for believing in me and investing in my dreams.

Acknowledgments

Above all, I thank David Autor and Heidi Williams for their encouragement and guidance. Anytime I radioed in and said, “Houston, we have a problem,” they immediately made themselves available and provided the exact support that I needed. David Autor is a true *mensch*. I am extremely grateful for David and Heidi’s advising and will be forever indebted to them. They are consummate academics and I deeply admire their dedication to doing impactful research on society’s most important topics. It was a privilege to have had them as advisers.

I thank my co-authors John Horton and Adam Kapelner without whom this dissertation could not have been written. John possesses unbounded intellectual curiosity and is one of the most creative people I have ever worked with. Our conversations throughout the years left an indelible mark on how I think about economics and online labor markets. I am also grateful for the doors he opened, for always having my best interests in mind, and for providing valuable advice during my career as a researcher and grad student. Adam has been both a co-author and a best friend. He is a kindred spirit and I am extremely fortunate to have had him as a lifelong intellectual companion. I am continually amazed by his spirituality and *NF* worldview. Additionally, he is a true wizard who possesses an extraordinary skill set and boldly applies it to bring his ideas into the world.

I am extremely grateful for having been part of the MIT Economics environment — there truly isn’t a better place for graduate training. The skills I developed here will serve me for the rest of my life. I was privileged to share the company of such amazingly talented classmates, TAs, and professors who taught me. I especially thank David Jiménez-Gomez for being a great friend and fellow student. He is a true philosopher/poet/scientist and I look forward to overcoming my own cost of thinking in order to appreciate his research on *the cost of thinking about thinking* and other topics.

Before MIT, there was the U of C. I first learned to love economics through the work of the late Gary Becker and others from the Chicago school of economics. More than any other intellectual approach, Chicago-style economics helps me see immense amounts of beauty in the world and understand the forces that guide human action. As a lecturer at MIT, I felt compelled to teach this perspective to my undergraduates and I hope that I inspired at least a few of them to appreciate economics.

I thank Bob Gibbons for being the greatest thesis adviser I never had. He was a tremendous teacher and enriched my understanding of incentives and organizations.

Additionally, I thank the many people from my personal life who supported me directly and

indirectly during grad school. I'm afraid to try to list them since I'll inevitably leave someone out who is truly important. However, I especially thank Johana for all of her love and support.

I thank oDesk and the unnamed firm for their efforts to create online labor markets that help make the world a more equitable place and unlock human potential. I am grateful to the long list of people there who helped me accomplish this research.

I gratefully acknowledge financial support from the George and Obie Schultz Fund and the NSF Graduate Research Fellowship Program. The NSF placed a great deal of faith in me and I hope that my work outside of academia brings about the broader impacts for which they supported me.

Finally, I thank Steven Levitt who has, without a doubt, been the most positive and persistent shock to my life. I will always remember the first day I started working for him and how he told me to view my RAship as, beyond anything else, an investment in my own human capital. I am honored that he saw potential within me and I appreciate all he has done to help me realize it. He has been a kingmaker to countless people and I am just one of many people who have benefited from his support. He is a model of what a mentor should be and he inspires me to pay it forward and be a mentor to others. He is also a tremendous institution builder; creator of human capital factories; and a wonderful asset to academia, the intelligent public, and the economics profession.

Contents

1	Market Congestion and Application Costs	15
1.1	Introduction	15
1.2	Experimental setting and design	18
1.2.1	The oDesk labor market	18
1.2.2	Experimental design	20
1.2.3	Balance checks	22
1.3	Evidence on applicant pool selection	22
1.3.1	Size of the applicant pool	23
1.3.2	Selection across applicant pools	23
1.4	Evidence on hiring behavior and employment outcomes	25
1.4.1	Hiring behavior	26
1.4.2	Employment outcomes	27
1.4.3	Heterogeneity in ordeal and information effects	31
1.5	Conclusion	32
2	Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets	51
2.1	Introduction	51
2.2	Mechanical Turk and its potential for field experimentation	54
2.2.1	General use by social scientists	54
2.2.2	Suitability for natural field experiments in Economics	55
2.3	Experimental Design	57
2.3.1	Subject recruitment	57
2.3.2	Description of experimental conditions	57
2.3.3	Task interface, incentive structure, and response variables	58
2.3.4	Hypotheses	59
2.4	Experimental Results and Discussion	60

2.4.1	Labor Participation Results: “Induced to work”	60
2.4.2	Quantity Results: Number of images labeled	61
2.4.3	Quality Results: Accuracy of labeling	62
2.4.4	Post Manipulation Check Results	63
2.5	Conclusion	64
2.6	Appendix	68
A	Detailed Experimental Design	68
3	Management and Measurement: Do More Attentive Supervisors Get Better Results?	75
3.1	Introduction	75
3.2	Setting: The firm and production environment	77
3.2.1	Company background and project description	77
3.2.2	The production process and staffing	78
3.2.3	Supervisor effort and its effect on correction rates	79
3.3	Data and summary statistics	80
3.3.1	The sample	80
3.3.2	Summary statistics	81
3.3.3	Effect of supervisor effort on correction rates for individual workers	84
3.4	Results	84
3.4.1	Results for worker behavior	85
3.4.2	Results for answer quality	86
3.4.3	Mechanisms for how supervision affects quality	89
3.5	Conclusion	92

List of Figures

1-1	Example job post: Interface workers see when deciding whether to apply to a job	33
1-2	Example job application: Interface workers use to submit applications	34
1-3	Experimental treatments	35
1-4	Flowchart of application process and the effect of the experiment	36
1-5	Flowchart of selection process and the effect of the experiment	37
1-6	Effect of treatment on number of applications	38
2-1	Main task portal for a subject in the meaningful treatment	65
2-2	The HIT as initially encountered on MTurk	68
2-3	The colorblindness test	69
2-4	Opening screen of training video	70
2-5	Examples of “cues of meaning”	70
2-6	Illustration of how to perform the task	71
2-7	Quiz after watching the training video (in the meaningful treatment)	72
2-8	Interface for labeling images	73
2-9	Landing page after completing a task	74
2-10	Post-task survey	74
3-1	Example question from a Q&A website	94
3-2	Flowchart of production process	95
3-3	Heterogeneity in supervisor effort and the average size their corrections	96
3-4	Visual IV: Supervisor effort, correction rate, and answer quality	97

List of Tables

1-1	List of questions used in treatment	39
1-2	Summary statistics for workers	40
1-3	Summary statistics for employers	41
1-4	Covariate balance: Employer characteristics by treatment cell	42
1-5	Applicant pool selection by treatment	43
1-6	Hiring outcomes	44
1-7	Employment outcomes: Job performance	45
1-8	Employment outcomes: Contract spend	46
1-9	Heterogeneity of ordeal and information effects: by job category, contract type, and employer experience	47
1-10	Applicant pool selection by treatment (additional variables)	48
1-11	Comparison of information effect controlling for worker quality index	49
2-1	Summary statistics for response variables and demographics by treatment and country	65
2-2	A heatmap illustration of our results	66
2-3	Main treatment effects on quantity of images	66
2-4	Main treatment effects on quality of images	67
3-1	Summary statistics: Questions and answers	98
3-2	Summary statistics: Worker behavior	99
3-3	Summary statistics: Supervisor behavior	100
3-4	Effect of predicted supervisor effort on corrections	101
3-5	Effect of supervisor effort on worker behavior	102
3-6	Effect of supervisor effort on answer quality	103
3-7	Effect of supervisor effort on additional measures of performance	104
3-8	Effect of supervisor effort on answer quality (by worker experience)	105

3-9	Decomposition of quality changes on corrected vs. uncorrected answers	106
3-10	IV estimates of the value of supervisor corrections	107
3-11	Variability by shift: Number of questions, answers, workers, and supervisors	108
3-12	Propensity of reviewers to make corrections	109
3-13	Worker exposure to supervisors: Fraction of question-answers reviewed by their top-5 supervisors	110

Chapter 1

Market Congestion and Application Costs

Abstract

We report the results of an experiment that increased job application costs in an online labor market. More specifically, we made it costlier to apply to jobs by adding required questions to job applications that were designed to elicit high-bandwidth information about workers. Our experimental design allows us to separate the effect of a costly ordeal vs. the role of information by randomizing whether employers see workers' answers. We find that our ordeal reduced the number of applicants by as much as 29% and reduced hires by as much as 3.6%. Overall, the applicant pool that underwent the ordeal had higher earnings and hourly wages, but not better past job performance. The ordeal also discouraged non-North American workers. We find no evidence that employers spent more when vacancies were filled, but some evidence that employer satisfaction improved. These improvements were the result of information provision rather than selection. Finally, we did not find any heterogeneity in outcomes across job category, contract types, or employer experience.

1.1 Introduction

Computer-mediation and digitization have dramatically lowered the cost of finding and applying to vacancies. Electronic job boards such as Monster.com and Careerbuilder.com have proliferated and workers can now search through thousands of job openings and easily submit hundreds of applications. Consequently, employers have access to more job candidates. However, it is not clear whether lower application costs have increased labor market efficiency or resulted in better matches (Kuhn and Mansour, 2013; Kuhn and Skuterud, 2004; Stevenson, 2008).

Why is this? One explanation is that the availability of applicants may not be the constrain-

ing factor. Rather, the true constraints may be high screening costs and the scarcity of “high-bandwidth” information about job candidates (e.g., data that “can only be verified through direct interactions such as interviews and repeated contact” (Autor, 2001)).¹ To provide an extreme example, if firms had only n_i interview slots and n_a applicants ($n_a > n_i$), increasing n_a would not help employers unless the candidates became more positively selected or employers could better decide whom to interview.

Another possible explanation is that very low application costs could reduce the signaling value of a job application. If employers have an imperfect screening technology and the applicant pool becomes adversely selected, employers may be worse off even with a larger candidate pool. In this situation, raising application costs could attract higher-quality candidates who are more likely to be genuinely interested in the position. However, the direction of selection is ultimately an empirical question.

In the simplest story, workers have priors on the expected match surplus and choose to submit an application whenever the expected surplus exceeds the cost of an application. If the expected match surplus correlates with the true match surplus, raising the application costs will increase the applicant quality. However, selection could go the other way. For example, if higher application costs took the form of a higher time cost, the best candidates may be discouraged if they have a higher opportunity cost of time.²

In this paper, we experimentally increase the cost of applying to jobs in an online labor market and examine the relationship between application costs and labor market outcomes.

Our first analysis examines to what extent adding questions induced applicant pool selection. Additionally, we designed the ordeal to elicit high-bandwidth information from the candidates: namely, by requiring candidates to answer questions that were added to randomly selected job posts.³ In order for us to measure the effect of information, we created two treatment arms with the same number of questions, but only in one of the arms did we show the answers to employers. Thus, the value of information can be identified by comparing the two treatments.

Our experimental sample includes oDesk employers who posted a job between November

¹ As noted by Autor (2001), the Internet does a good job transmitting low-bandwidth data (i.e., “objectively verifiable information such as education, credentials, experience, and salaries,”) and helps employers and workers find each other more easily, but may fail to produce the kind of high-bandwidth data that facilitates better matching.

²As noted by (Alatas et al., 2012), the degree of selection doesn’t have to be non-monotonic and the theoretical relationship is ambiguous once plausible real-world features are added to the model. Therefore, it is essential to treat the relationship between ordeal size and selection as an empirical question.

³Example questions include: “Why are you a good fit for this job?” “What part of this project most appeals to you?” table 1-1 shows the full list of questions.

21, 2012, and May 1, 2013. Employers were assigned to the control group (72%) or one of four treatments that required questions.⁴ In three of the treatments, candidates answered one (7%), two (7%), or five (7%) questions and their answers were shown to employers. In the final treatment (7%), candidates answered five questions, but their answers were not shown to employers. By comparing this last treatment to the other five-question treatment where employers saw answers, we measure the effect of eliciting more information, which we call the “information effect” and by comparing it with the control, we measure the effect of applicant pool selection, which we call the “ordeal effect.”

We find that the primary effect of the intervention was to reduce the size and change the composition of the applicant pool. The applicant pool decreased by nearly thirty percent and the workers who underwent the ordeal were more positively selected. Vacancies were 3.6% less likely to fill, but this is a small decrease given the large overall reduction in candidates. The matches that were formed did not have a higher wage bill and the candidates who were hired did not differ on observable characteristics. However, we find some evidence that job performance among the matches improved by 5.5% and that these effects were concentrated in the treatments that acted through an information effect. Importantly, we caution that the results for wage bill and job performance cannot be easily interpreted since they are *conditional-on-positive* outcomes and may be biased if the types of vacancies filled or quality of workers differ across treatment cells. Finally, we find no evidence of heterogeneous responses across job type, contract type, or employer experience.

This experiment is motivated by several literatures. The first literature is that of ordeals and self-targeting. Governments frequently face the problem of targeting transfer programs to the neediest recipients. With imperfect information, governments may sacrifice productive efficiency by imposing “tedious administrative procedures” and “ordeals” that result in dead-weight loss in order to improve target efficiency (Nichols and Zeckhauser, 1982). Examples include “workfare” programs that require recipients to perform (often unproductive) work in order to receive benefits (Besley and Coate, 1992), conditional cash transfer programs that require recipients to travel greater distances (Alatas et al., 2012), and even food aid programs that deliberately provide lower-quality goods that carry social stigma (Tuck and Lindert, 1996).

The second major literature is that of labor market congestion and preference signals. In congested markets, low application costs encourage job seekers to apply to an “excessive” number of positions, including ones they likely would not accept. Employers are often faced with

⁴After being assigned to a group, all subsequent job posts by that employer were treated according to the initial assignment.

more candidates than they have interview slots and worry about wasting interview slots on unattainable candidates (Coles et al., 2010a). As a result, truly interested candidates who appear overqualified may not receive interviews even if firms would want to hire them. To overcome congestion, several markets have implemented mechanisms that allow candidates to indicate top choices (a.k.a., “preference signals”). For example, the AEA job market for new economists gives participants a fixed number of signals that applicants can send to employers to indicate special interest (Coles et al., 2010b).⁵ Similarly, a Korean online dating site allowed participants to send a limited number of virtual roses to indicate special interest (Lee et al., 2011). In both cases, introducing preference signals improved matching, especially for candidates who seemed overqualified.

This paper proceeds as follows. Section 1.2 describes the oDesk labor market and our experimental design. Section 1.3 analyzes how the intervention influenced the size and composition of the applicant pool. Section 1.4 examines hiring and employment outcomes and section 3.5 concludes.

1.2 Experimental setting and design

1.2.1 The oDesk labor market

During the last ten years, a number of online labor markets have emerged. We conducted our experiment on the largest of these online labor markets, oDesk.

The processes on oDesk mirror those in traditional labor markets: employers post vacancies, conduct recruitment, and screen workers, while Workers search through jobs and submit applications. Figure 1-1 shows an example job application. As of October 2012, more than 495,000 employers and 2.5 million workers have created profiles on oDesk. In 2012, 1.5 million jobs were posted, workers logged over 35 million hours, and employers paid more than \$360 million in wages.⁶

oDesk provides important services to the marketplace including: maintaining job listings, arbitrating disputes, and creating reputation mechanisms. When a job is posted, oDesk also uses a “recommender system” to algorithmically identify qualified workers and suggest them

⁵Avery and Levin (2010) study the effect of allowing candidates to indicate special preference for colleges by applying through early admissions policies.

⁶Due to competitive concerns, we do not report overall marketplace statistics that are not publicly available. We are able to report these statistics since they are derived from a press kit that was distributed to newspapers for citation.

to employers. Once workers are hired, oDesk handles payment processing and provides worker management tools such as time-tracking software that helps employers monitor workers by taking screenshots of their computer.⁷

Types of work

There are two main job categories: information technology (IT) and knowledge process outsourcing (KPO), and two main contract types: fixed-price (FP) and hourly (HR).

IT jobs generally require higher-skilled workers than KPO jobs. The most common IT jobs are web programming and web design, while the most common KPO jobs are data entry and web research. Although KPO jobs account for a larger number of hours, they are lower-paid and the majority of total wages come from higher-skilled IT work.

In a fixed-price contract, employers pay based on a final deliverable, whereas in hourly contracts workers are paid as they log hours. Hourly contracts tend to be longer duration and higher billing. Although, ~ 60% of contracts are hourly, they make up nearly ~ 85% of total billings.

Worker and employer characteristics

The majority of workers and employers in our experiment were new to the oDesk platform. 58.5% of employers and 25.3% of workers had been on oDesk less than two weeks. Employers and workers are predominantly male, 78.9% and 69.2%. The vast majority (83.1%) of employers come from developed countries and the most represented regions are North America (57.5%), Western Europe (15.8%), and Australasia (9.8%). Workers, on the other hand, are predominantly drawn from South Asia (44.1%), East Asia (22.8%), and Western Europe (7.6%).

Only 40.1% of employers had posted a job in the past and only 30.6% of employers had made past hires. Of those, the median employer hired 5 workers and the median total spend was \$548 (the respective means are 11.0 hires and \$2,764 past spending). For workers who had been hired on oDesk, the median worker had been hired for 5 jobs, worked 155 hours, and earned \$715. The respective means are 12.5 jobs, 657 hours, and \$4,537 in earnings. For hourly contracts, workers earned a mean of \$7.92/hr (median = \$5.15/hr). For fixed-price contracts, the average contract was \$78.38 (median = 18.00).

Workers may also specify outside-of-oDesk characteristics such as education, years of work history, and past employers. 56.0% of workers have more than 5 years of non-oDesk work experience and 75.3% of workers are college educated. Workers also specify a desired hourly wage (their “profile wage”), which averages \$7.92/hr. Although there are workers of all skill levels,

⁷This software, known as the “Work Diary,” lowers monitoring costs and helps facilitate time-and-materials contracts, which may be more efficient depending on the nature of the work.

over a quarter of workers indicate in their profiles that they are willing to work for \$2.22/hr or less and the median worker only expects \$5.15/hr.

Tables 1-2 and 1-3 provide further summary statistics for workers and employers, respectively.⁸

1.2.2 Experimental design

This section is divided into two subsections. In the first, we describe the treatment cells and randomization. In the second, we explain how the treatments identify the causal channels through which the intervention may have operated.

Overview of treatments

Our experiment took place between November 21, 2012, until May 1, 2013, during which time 90,336 employers posted 239,009 job posts and received more than two million applications from 222,671 workers.⁹

Randomization was done at the employer level. Upon posting a job, each *employer* was assigned either to the control or one of four treatment cells. Employers' assignments determined the question requirements of the initial and all subsequent job posts. All workers who applied to a given vacancy answered the same questions.

The four treatment cells varied along two dimensions:

- **Number of required questions:** Applicants have to answer one, two, or five questions
- **Information provision:** Employers were randomly shown answers

Prospective candidates who viewed treated jobs only saw the number of required questions and were not told whether employers would see the answers.¹⁰ The job applications appeared identical except that the treated posts had space for additional questions (see figure 1-2).

The three treatments where employers saw answers were denoted: *Q1-Ans*, *Q2-Ans*, and *Q5-Ans*. The fourth treatment, which contained five questions that were not shown to employers,

⁸Employers characteristics are measured at the time they post a job and enter our experiment. Worker characteristics are measured at the time they first submit an application.

⁹In addition to the 239,009 jobs, which were publicly visible to the marketplace, there were "private" and "invalid" job posts, both of which we exclude. We exclude private job posts because they were not visible to the marketplace and workers only found out about them if they were invited by the employer (who often already knew the worker). Since our experiment is about applicant behavior, private posts are not informative. Invalid job posts are excluded because they do not represent actual job posts; those posts were removed by oDesk moderators for violating terms of service (e.g., spam, advertising).

¹⁰We expect that workers assumed that their answers would be seen by employers.

was denoted *Q5-NoAns*. 28% of job posts were assigned to one of the four treatments (with an equal 7% split). Figure 1-3 shows a matrix summarizing the treatments.

Figure 1-4 provides a flowchart of the stages of the experiment. The basic stages are: 1) Workers find or are invited to a job post; 2) workers see the question requirement and decide whether to apply; 3) if an applicant answered questions, employers are shown the questions (depending on treatment); and 4) employers decide whether to make a hire.

Identifying causal channels: ordeal vs. information effect

This section describes how our test cells disentangle the effects of: (1) selection that was induced by the ordeal and (2) revelation of high-bandwidth information that was elicited by the required questions. Section 1.3 presents evidence on applicant pool selection and section 1.4 examines hiring and employment outcomes.

In particular, we analyze:

1. **Applicant pool selection:** how the questions changed the size and composition of the applicant pool
2. **Pure ordeal effect:** how applicant pool selection affected hiring and employment outcomes due to selection caused by the ordeal *holding information constant*
3. **Pure information effect:** how providing employers with answers affected hiring and employment outcomes *holding applicant pool selection constant*

Exercise 1 is descriptive and describes how the ordeal influenced workers' application decisions and the resulting applicant pool selection. We accomplish this by comparing worker characteristics for each ordeal size.

Exercise 2 measures the pure ordeal effect and how applicant pool selection affected hiring and employment outcomes. More specifically, we identify this effect by comparing *Q5-NoAns* to the control group. Both treatments provide identical (i.e., zero) information to employers, but the candidates in *Q5-NoAns* are selected since they had to undergo the ordeal.

Exercise 3 measures the pure information effect and how providing answers to employers affected hiring and employment outcomes. More specifically, we identify this effect by comparing *Q5-Ans* to *Q5-NoAns*. Both treatments required candidates to undergo an identical ordeal (answering five questions), but *Q5-Ans* showed information to employers.

The flowchart in figure 1-5 summarizes the stages of the application process and the points at which the ordeal and information effects are identified.

1.2.3 Balance checks

In order to ensure that our randomization was successful, we perform two balance checks: 1) whether employer characteristics were balanced across experimental treatments, and 2) whether worker characteristics were balanced in the Q5-Ans and Q5-NoAns treatments.

Table 1-4 tests balance by regressing each employer characteristic on indicators for treatment cell. Additionally, the row “F-test (p-value)” tests the joint hypothesis that all treatment dummies are equal to zero. We fail to reject the null for any variable. Of eight covariates and four treatment dummies tested only one (total # of hires) is significantly different at the 5% level.

Next, we check whether worker characteristics are balanced in the Q5-Ans and Q5-NoAns treatments. Since both treatments appeared identical to candidates (i.e., workers saw five questions, but didn’t know whether the answers would be seen), we expect that worker characteristics should be the same for both treatments. The row “Delta Q5” of table 1-5 reports p-values from a test of whether the treatment effects for Q5-Ans and Q5-NoAns are equal; we fail to reject equality for every worker characteristic. To further verify balance, we use data from Q5-Ans and Q5-NoAns and regress an indicator for “in Q5-Ans” on *all* worker characteristics. The resulting F-test has a p-value of 0.215, suggesting that we are balanced.

1.3 Evidence on applicant pool selection

This section examines the impact that the ordeal had on the applicant pool. In the first part, we analyze the overall reduction in the size of the applicant pool. In the second part, we analyze whether the ordeal induces selection among the applicant pool.

For the first analysis, where each unit of observation is a job opening (equivalently an employer), we do not cluster standard errors. For the second analysis, where the unit of observation is an individual job application, we cluster standard errors on the job opening. Throughout our analyses, we estimate effects using only the *first* job opening posted by an employer after being allocated to a treatment cell. This simplifies interpretation since we do not have to worry about how the randomization may have affected the creation of future posts.

1.3.1 Size of the applicant pool

We estimate changes in the size of the applicant pool with the following equations:

$$\begin{aligned} N_i &= \alpha + \beta \cdot \text{Q1-Ans}_i + \gamma \cdot \text{Q2-Ans}_i + \delta \cdot \text{Q5-Ans}_i + \eta \cdot \text{Q5-NoAns}_i + \varepsilon_i \\ N_i &= \alpha + \beta_1 \cdot \text{ONE}_i + \beta_2 \cdot \text{TWO}_i + \beta_5 \cdot \text{FIVE}_i + \varepsilon_i \end{aligned} \tag{1.1}$$

where N_i is the number of applications for opening i . In the first equation, we let β , γ , δ , and η represent the treatment effects. Additionally, we modify the first equation to estimate separate coefficients based on the number of questions in each treatment.¹¹ This eases discussion of results and is justified since candidates responded identically to the two Q5 treatments (see section 1.2.3).

The additional questions substantially reduced the number of applicants. Moreover, there was a monotonic relationship between the size of the applicant pool and the size of the ordeal. From a base of 25.7 applicants in the control group, adding just one question decreased the pool by 1.7 applicants, adding two questions reduced it by 4.1 and adding five questions reduced it by 7.3, a nearly 30% decrease (see table 1-6). Additionally, these reductions occurred across job categories and contract types. These results are all significant at the 1% level. Figure 1-6 graphs the mean and standard error of the number of applications across job categories and contract types.

Holding the quality of the applicant pool constant, employers should be weakly worse off with fewer candidates. Therefore, for the ordeal to help employers, it would need to induce positive selection.

1.3.2 Selection across applicant pools

This subsection examines whether the ordeal induced positive selection and how worker characteristics changed. Our goal is to determine whether the ordeal resulted in positive selection. To do this, we examine the following worker characteristics: 1) demographics, experience, and earnings; 2) past job performance; and 3) an overall worker quality index that combines all characteristics.

¹¹Thus, we replace β with β_1 and γ with β_2 to represent the one- and two-question treatments and combine Q5-Ans and Q5-NoAns (the five-question treatments) to estimate β_5 .

We estimate differences in applicant pool characteristics by the following equations:

$$\begin{aligned} X'_{ij} &= \alpha + \beta \cdot \text{Q1-Ans}_{ij} + \gamma \cdot \text{Q2-Ans}_{ij} + \delta \cdot \text{Q5-Ans}_{ij} + \eta \cdot \text{Q5-NoAns}_{ij} + \varepsilon_{ij} \\ X'_{ij} &= \alpha + \beta_1 \cdot \text{ONE}_{ij} + \beta_2 \cdot \text{TWO}_{ij} + \beta_5 \cdot \text{FIVE}_{ij} + \varepsilon_{ij} \end{aligned} \quad (1.2)$$

where X'_{ij} is a vector of characteristics for worker i who applied to opening j . As was done in Equation 1.1, the second line pools the two Q5 treatments. Additionally, we cluster standard errors at the job opening level¹² to avoid artificially inflating the size of our data set by using individual job applications.

Demographics, experience, earnings, and past performance

We report our main applicant pool selection results in table 1-5. Each column shows a regression of a particular worker characteristic on dummies for the size of the ordeal (i.e., number of questions). There was positive selection for most of these variables, although the effect was not statistically significant in the one-question treatment. We also find that the larger, five-question ordeals induced more selection (see the row “ordeal size”). Appendix table 1-10 present similar results for additional characteristics.

Candidates who underwent the ordeal had substantially more experience. Candidates who had answered two questions had been on oDesk 2.5% longer and candidates who had answered five questions had been on oDesk 4.1% longer (column 2). They were slightly more likely to have been hired in the past (between 1.2 and 3.6 percentage points), and, among those who had been hired, candidates had between 3.5% and 6.0% more past hires (columns 4 and 5). Candidates did not spend a significantly different number of hours working, but those who worked had between 8.9% and 12.5% higher total earnings (column 7).

Workers’ higher earnings in the ordeal treatments were driven by higher wages on both hourly and FP contracts (columns 1 and 2 of 1-10), suggesting that they are more skilled. Additionally, their salary expectations (as reported by their profile wage) were between 6.1% and 8.1% higher.

Column 9 of table 1-5 shows that the ordeal increased the proportion of North American candidates in the two- and five-question ordeal groups by as much as 8.6%, perhaps because workers outside of North America were more likely to be non-native English speakers and were more deterred by the questions.

¹² This is equivalent to the employer level since our analysis is restricted to an employer’s first job post.

Past job performance, gender, age, non-oDesk work experience, and educational attainment were not different across treatments (see appendix table 1-10).

Overall quality index

Finally, rather than look at variable-by-variable differences in worker quality, we create a single quality metric for each worker by estimating each worker’s likelihood of being hired in the marketplace. More specifically, we regress a dummy for “hired” on all worker characteristics for all applicants in the control group.¹³ Using the estimated model, we calculate fitted values for workers in all groups and report them in column 1 of table 1-5.

The average probability that an applicant from the control group was hired is 1.6%. Although this probability did not increase substantially, it was approximately 1.6% higher in the two- and five-question treatments relative to the control. This difference is much smaller than the differences in the aforementioned characteristics. One explanation is that greater experience is not indicative of worker quality. Another explanation is that our worker quality index is not a good proxy for quality (*psuedo-R*² = 0.0258).

1.4 Evidence on hiring behavior and employment outcomes

The past section examined applicant pool selection, which only depended on workers’ response to the ordeal. This section analyzes outcomes that are also the result of *employers’ decisions*, namely: 1) hiring behavior (e.g., vacancy fill rate, number of interviews) and 2) employment outcomes related to a hire (e.g., amount spent on a contract).

We use the following equation to estimate differences in these outcomes across treatments:

$$Y_i = \alpha + \beta \cdot Q1\text{-Ans}_i + \gamma \cdot Q2\text{-Ans}_i + \delta \cdot Q5\text{-Ans}_i + \eta \cdot Q5\text{-NoAns}_i + \epsilon_i \quad (1.3)$$

Note: Uses data from all treatments

where Y_i measures a hiring or employment outcome for opening i and β , γ , η , and δ measure the treatment effects. Since our unit of observation is a job opening, we do not cluster standard errors.

¹³To handle missing values, we use indicator variables to indicate missingness. For categorical variables like gender, we create a dummy variable for “gender = missing”. For quantitative variables such as log earnings, we create a dummy variable for “log earnings = missing” and substitute the mean value of non-missing rows into the missing rows.

A key feature of our experimental design is that we can disentangle the effect of imposing an ordeal (the **pure ordeal effect**) vs. the effect of providing information to employers (the **pure information effect**). For clarity, the below equations show how we estimate these effects:¹⁴

$$Y_i = \alpha + \delta^{ORD} \cdot \underbrace{Q5\text{-NoAns}_i}_{ORDEAL} + \varepsilon_i \quad (\text{ordeal effect}) \quad (1.4a)$$

Note: Uses data from control and Q5-NoAns

$$Y_i = \alpha + \delta^{INFO} \cdot \underbrace{Q5\text{-Ans}_i}_{INFORMATION} + \varepsilon_i \quad (\text{information effect}) \quad (1.4b)$$

Note: Uses data from Q5-Ans and Q5-NoAns

Equation 1.4a estimates the pure ordeal effect (η^{ORD}) by measuring the difference between Q5-NoAns and the control. Equation 1.4b estimates the pure information effect (δ^{INFO}) by measuring the difference between Q5-Ans and Q5-NoAns.

1.4.1 Hiring behavior

This section examines how hiring and recruiting behavior were affected. The primary outcomes we consider are: 1) the number of invitations employers send; 2) the fill rate of vacancies; 3) whether vacancies were more likely to be filled from applicants vs. invitations; 4) the wage candidates are hired at; and 5) the quality of applicants who are hired.

Employers primarily made hires from the unsolicited applicant pool. Although the overall hire rate of unsolicited applicants was lower (3.8% vs. 12.7%), the unsolicited candidate pool was larger, and 80.9% of positions were filled from there. Only 41.3% of employers made any invitations and of those who made invitations, 79.9% made five or fewer and 43.0% made just one invitation.

Despite the nearly thirty percent reduction in applicants, employers did not recruit more heavily on either the intensive or extensive margins for any groups (see columns 3 and 4 of table 1-6). We also fail to find any effects when estimating a quasi-maximum likelihood Poisson model.¹⁵ Even without additional recruiting, the probability of making a hire in the one- or

¹⁴The coefficients δ^{ORD} and δ^{INFO} corresponds to δ and $\delta - \eta$, respectively, from Equation 1.3.

¹⁵Chapter 20 of Cameron and Trivedi (2005) provides a thorough discussion of Poisson regression models and their use.

two-question treatment group remained constant and only decreased by 3.6% (a decrease of 1.4 percentage points from a base of 39.3, $p < 0.05$) among the five-question treatments (column 5 of table 1-6).¹⁶

Next, we analyze whether the quality of workers who were hired differed across treatments. Table 1-6 shows that there were no differences according to the worker quality index (column 7) or profile wages (column 8). Additionally, the wages they received were not higher for fixed or hourly workers (columns 9 and 10).

Finally, we test whether showing answers to employers affected hiring outcomes *after* controlling for selection. To measure this effect, we report estimates of the difference in treatment effects between Q5-Ans and Q5-NoAns (Q5-Ans minus the Q5-NoAns) and report results in row “Information effect” of table 1-6. We find that not a single outcomes differed due to showing information, which suggests that the changes in hiring outcomes was due to selection, rather than information.

1.4.2 Employment outcomes

This section analyzes the success of a match once a worker was hired to a vacancy using: 1) worker job performance and 2) wage bill (i.e., the amount spent on a contract).

Before proceeding to the results, we caution that both of these outcomes are “conditional-on-positive” outcomes and may be hard to interpret. Outcomes for job performance are only defined if a vacancy is filled and, if the types of vacancies that are filled or the quality of workers hired differ among the treatments, our treatment effects would be confounded with selection. Given that the fill rate of vacancies declined (see section 1-6), this is a genuine concern for how we interpret the results. Likewise, the same conditional-on-positives applies for the wage bill.

Job performance

Overall, we find that the matches formed from workers who underwent the ordeal did not perform better on contracts. However, when using our preferred metric for successful performance, we find some evidence that showing employers answers results in better matches (i.e., that the effect occurs through information rather than selection).

Our primary metrics are: the rating an employer gave the worker and whether the employer

¹⁶Given the overall reduction, the probability that an opening was filled from the unsolicited applicant pool also decreased (column 6 of table 1-6). As a necessary consequence, the hiring rate increased due to the fact that the ordeal reduced the applicant pool by nearly thirty percent and the reduction in hiring was much smaller.

told oDesk that a contract ended successfully.¹⁷

The first metric, the rating an employer gave the worker, is visible to the marketplace. oDesk workers are rated on a five-point scale along six dimensions (e.g., communication, availability, and work quality). However, this metric has very little variation and tends to be inflated because employers are reluctant to give poor reviews out of sympathy towards the worker.¹⁸

The second metric, whether the employer said a contract ended successfully has much more variation and, because it's reported privately, is less likely to suffer from bias. While workers received perfect five-star ratings for 61.6% of contracts, only 50.6% of contracts were reported as successful. After the time of this study, oDesk redesigned its feedback collection process to rely more on privately reported feedback, which supports the notion that the second metric is a better indicator of job performance.¹⁹

Table 1-7 reports results for job performance. The most notable result is that contracts in the pure information treatment are more likely to be successful, but that contracts in the pure selection treatment are not (column 1). In particular, Q5-Ans, contracts ended successfully 2.8 percentage points more often (a 5.5% increase) and the treatment effect of Q5-Ans minus Q5-NoAns was 3.4% higher ($p < 0.05$). There was no corresponding increase in any of the one- or two-question treatments.

This suggests that having access to the workers who were more positively selected (in the Q5-NoAns treatment) was not enough to make contracts more successful and that outcomes only improved when employers saw the answers.

Next, column 2 shows the treatment effects on five-star feedback (after normalizing by the mean and standard deviation). None of the treatment effects are significant at the 5% level and we can reject effects larger than approximately 0.1 standard deviations.

The remaining columns of table 1-7 analyze other measures of success including whether: 1) employers paid a bonus on an hourly contract²⁰, 2) employers paid more than the contracted amount for a FP contract, or 3) employers requested a refund. We find no statistically significant

¹⁷The options were: "Job completed successfully," "Job completed unsuccessfully," and "Job cancelled or postponed".

¹⁸An oDesk survey of employers reported that many of them felt pressure to leave higher-than-desired ratings. The primary reason employers reported leaving higher ratings was sympathy towards the worker. The next most common reason was fear of retaliation from the worker (e.g., that the worker would rate them badly or potentially sabotage work that had been done).

¹⁹In the redesign process, oDesk frequently used statistics like we report above as a rationale for increasing the amount of privately collected feedback.

²⁰We don't analyze bonuses on fixed-price contracts since oDesk does not distinguish well between a milestone payment (representing ordinary payment) and a bonus (representing superior work). Instead, we proxy for success of a FP contract by whether payments exceeded the budget (measure 2).

differences in any of these additional measures.

Total wage bill of contracts

As a proxy for contract success, we examine the total amount spent (the wage bill) during the first 150²¹ days. We conclude that there is very little evidence that the wage bill is higher in any of our treatments. However, even if we did find differences, it's not clear that the total wage bill is a good proxy for surplus except under very narrow assumptions. The wage bill would be a good proxy if we assumed that: 1) firms use labor in fixed proportions, 2) each unit produced earns positive per-unit profits for the firm, and 3) that workers' wages exceed their outside option. However, under different assumptions, the wage bill would be a poor proxy for surplus. For example, if there were perfect competition and if firms paid workers their marginal product, a larger wage bill would not result in higher employer surplus since per-unit profits would be zero.²²

Even if we believed that the total wage bill were a good proxy for surplus, the distribution of the variable would make it complicated to interpret. Namely, only about 40 percent of job posts are filled and the majority of values are zero. Additionally, the variable is highly skewed, which limits the ability to use the untransformed values.²³ Consequently, we estimate our models using several specifications: 1) regressing a binary indicator for whether a job opening had any spending, 2) an ordinary linear specification, 3) log and log(1+x) transformations of the dependent variable, 4) quantile regressions, 5) regressing a binary indicator for an above-median wage bill, 6) a quantile regression, and 7) a quasi-maximum likelihood Poisson regression.²⁴ We report all of these estimates in table 1-8.

Remarkably, we find no statistically significant treatment effects in *any* of these models. In the simplest estimation strategy, we regress an indicator for whether an opening logs *any* spending (column 1). All of these coefficients are negative, which is unsurprisingly given our finding in section 1.3.1 that the probability of filling a position decreases by more than three percent. Based on the confidence intervals for these estimates, we can generally reject that reduction is larger than about two percent. We also fail to find significant differences in the

²¹The majority of contracts end much sooner than that and the results do not substantially change when using 30- or 90-day intervals.

²²Likewise, if hours go up and wages remain constant, workers would not be better off and might simply end up working greater hours.

²³In other oDesk experiments, the most effective treatment would change if you excluded just a handful of the largest employers.

²⁴Chapter 20 of Cameron and Trivedi (2005) provides a thorough discussion of Poisson regression models and their use.

linear model (column 2).

Since our dependent variable is highly skilled, we estimate a log transformed model (column 3). In this model, the confidence intervals are much wider and, depending on the treatment, we are unable to reject effects as large as an increase of about ten percent and as small as a decrease of about ten percent. The coefficient that is closest to statistically significant ($p = 0.111$) indicates a 6.1% increase in spend for the Q5-Ans group. However, a major limitation of the log specification is we are forced to drop the majority of observations (because they are zeros). This introduces a potentially serious conditional-on-positive problem, especially since the fill rate of vacancy declined by three to four percent and the composition of openings that are filled may have changed as a result of the treatment.

Finally, we estimate additional specifications that are robust to missing values and skewed dependent variables: an indicator for whether an opening had spend above the category-specific median²⁵ (column 5), a quantile regression model using log spend (column 6) and a quasi-maximum likelihood Poisson regression (column 7). In none of these specifications do we find any effects that are significant at the 5% level. One potentially noteworthy results, is that the coefficient on the Q5-Ans treatment has a log quantile treatment effect (at the 90th quantile) of 0.137, which would indicate that providing information improved outcomes at the higher quantiles. However, the coefficient is not statistically significant ($p = 0.082$) and not robust to other specifications.²⁶

Finally, we investigate whether the answers helped employers learn valuable information about candidates beyond what was observable. In order to do this, we estimate the information effect *controlling for* the quality of the person hired, as measured by our quality index (see section 1.3.2), and we compare the estimates of the information effect with and without controls. In table 1-11, the “Unadjusted” row reports the ordinary information effect and the “Controlling for quality index” row reports how the estimates change after controlling for worker quality. The point estimates are virtually unchanged after adding controls, suggesting that employers do not learn additional information from the answers; however, the difference between the estimates is very imprecisely estimated. This is not surprising given that the quality index explains so little variation in quality ($psuedo-R^2 = 0.0258$). In conclusion, we consider this to be an extremely

²⁵This strategy avoids an issue created by the heterogeneity in spending across different categories. For example, if one category, such as software development, had substantially higher spend than all others and represented a large portion of the marketplace, measuring a treatment effect at the upper tail would only be informative for the software development category.

²⁶Although not included, we perform sensitivity checks using log and levels specifications at other quantiles and fail to find statistically significant effects. In almost all cases, when we graphed the quantile treatment effects for different treatments, we can draw a zero-line that is fully contained within the 95% confidence interval.

low-powered test and do not draw any conclusions from it.

1.4.3 Heterogeneity in ordeal and information effects

This section examines whether the ordeal effects or information effects differ across subsamples of the data. We use data from the control, Q5-Ans, and Q5-NoAns treatment cells to examine whether there is heterogeneity across the following dimensions:

- **Job category:** IT vs. KPO
- **Contract type:** Fixed-price vs. hourly
- **Employer experience:** Zero hires, below-median # hires, and above-median # of hires

and we estimate heterogeneous treatment effects for the following outcomes: 1) whether a vacancy is filled, 2) job performance, and 3) log total spend.

Table 1-9 reports estimates of the ordeal effect and information effect for the entire sample and within different subsamples. The columns in the table report the coefficients for the ordeal effect and information effect for each outcome. The first row presents estimates for the entire sample.²⁷ The following three panels present results for job category, contract type, and employer experience. Employer experience divides employers into three groups: 1) high experience, 2) low experience, and 3) no experience. Employers in groups one and two had made past hires and are divided into those who made above or below the number of hires. Employers in group three had never made a hire.

We find very little evidence of heterogeneous treatment effects for any of the outcomes. In several cases, the coefficients for the ordeal or information effect are statistically significant within one subsample, but not another. For example, the ordeal effect for fill rates (column 1) is statistically significant at the 5% level within the KPO, but not the IT category. However, the difference between these ordeal effects is not statistically significant.

Of the five cases where the ordeal or information effect is statistically significant in one group, but not others, the only case where the difference between the effects is close to statistically significant is for the ordeal effect for fixed price vs. hourly ($p = .060$).²⁸

²⁷These are the same estimates as those reported in tables 1-6, 1-7, and 1-8. Each ordeal effect is equivalent to the coefficient on Q5-NoAns and each information effect is the difference between the Q5-Ans and Q5-NoAns treatments.

²⁸For fill rates, the ordeal effect lowered the fill rate by a statistically significant amount for KPO, but not IT jobs. For the probability that a contract ended successfully, the information effect raised the fill rate by a statistically significant amount for KPO jobs, hourly jobs, and for employers without experience. However, none of these were significant.

1.5 Conclusion

Our paper analyzes an intervention in an online labor market where we increased the application costs by designing an ordeal that required applicants to answer additional questions that we added to their job applications. In our treatments, we also randomly showed their answers to employers which allows us to measure the effect of the information on employers' decisions. Consequently, we analyze the extent to which applicant pool selection vs. information provision affects outcomes.

We find that the primary impact of the intervention was to reduce the size of the applicant pool and change its composition. The applicant pool decreased by nearly 30 percent and the workers who underwent the ordeal were more positively selected. Vacancies were 3.6% less likely to be fill, but this represents a relative small decrease given the large overall reduction in candidates.

We find some evidence that the workers who filled positions had higher job performance, although the total spending on contracts did not increase. We also find that the largest increases in job performance occurred in Q5-Ans and that they were significantly different than the treatment effect in Q5-NoAns, which suggests that the improvement was due to information revelation rather than selection. However, as we cautioned before, job performance and total spending are both conditional-on-positive outcomes and may be difficult to interpret. We find no evidence of heterogeneous treatment effects for different employer types.

Given that there were only small improvements to outcomes, the intervention may have been more costly to candidates relative to the improvements in terms of the improved hiring and employment outcomes. Nevertheless, we caution that many of our outcomes are crude and that future interventions may be more successful if they are able to elicit higher quality information from candidates

Figure 1-1: Example job post: Interface workers see when deciding whether to apply to a job

The screenshot shows a web interface for job posting. At the top is a navigation bar with tabs: Recruit, Manage My Team, Find Work (active), My Jobs, Wallet, Reports, and Messages. Below this is a secondary navigation bar with links: Find Jobs, Job Applications, Profile, Staffing Console, and Tests. On the right of this bar are search and filter controls.

The main content area features a job title "Some Data Entry Fields" with a description: "Fixed Price Project - Est Budget \$300.00 - Posted 1 hour ago". A yellow callout box highlights the job title/description area. To the right, there are links for "Flag as inappropriate" and "Previous 12 of 2433 Next".

A prominent green button labeled "Apply to this Job" is enclosed in a dashed black box. Below it is a "Job Overview" section with the following details:

Type:	Fixed Price
Budget:	\$300.00
Posted:	April 29, 2013
Planned Start:	Immediately
Delivery Date:	May 1, 2013
Visibility:	Public
Category:	Administrative Support
Sub-Category:	Data Entry

On the left side, there are sections for "Job Description" (with a blurred text area), "Preferred Qualifications" (Contractor Type: Independent Contractors Only), and "Client Activity on this Job" (Last Viewed: 7, Applicants: 16, Interviewing: 0). A yellow callout box highlights the text: "Our treatment: Adds 1, 2 or 5 questions". Below this, a red-bordered box contains the text: "Number of questions to answer to apply: 5".

Notes: Before applying to a job, subjects in Q1-Ans, Q2-Ans, Q5-Ans, and Q5-NoAns see the number of questions required to apply (at bottom left).

Figure 1-2: Example job application: Interface workers use to submit applications

The screenshot shows a web interface for applying to a job. The main heading is "Apply to Job". Below it, the job title is "Marketing Expert to look over our site and suggest a business strategy." The description follows, stating it's a paid trial for a marketing expert. There are several highlighted areas: a yellow box for the job title and description, a yellow box for the wage bid section, a blue box for the cover letter, and a dashed red box for additional questions. The wage bid section includes a table with columns for "Paid to You", "+ 10% oDesk Fee", and "Charged to client", all showing "0.00 /hr". The "Submit application" button is highlighted in yellow.

Apply to Job

Job Posting **Marketing Expert to look over our site and suggest a business strategy.**
This is a paid trial for a marketing expert who will look over our site and suggest a master plan for us to proceed.
Candidates who provide us with a comprehensive plane will be hired for a long term assignment.
Do not apply if you are not and exp... more
[View job posting](#)

* Propose Terms: Propose an hourly rate of:

Paid to You: \$	0.00 /hr
+ 10% oDesk Fee: \$	0.00 /hr
Charged to client: \$	0.00 /hr

* Cover Letter:

Free-form cover letter

Appears for treatment cells

Additional questions (if applicable)

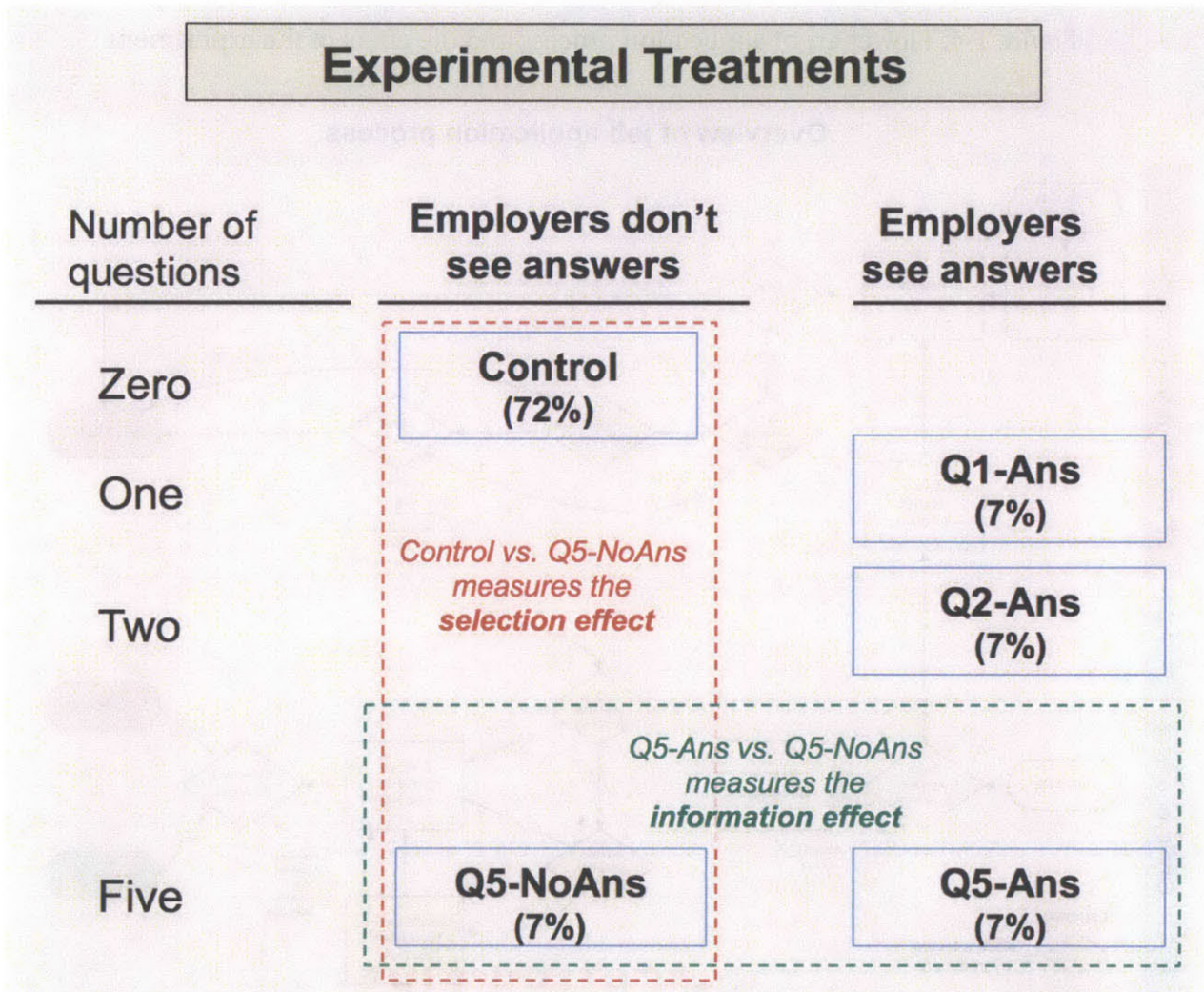
Attachment: no file selected
File size should be less than 5MB. Include work samples or other documents to support your application. Do not attach your résumé — your oDesk profile is automatically forwarded to the client with your application.

* Agree to Terms: By submitting my bid on this Contract, I agree to the terms and conditions of the oDesk User Agreement and incorporated Policies.

Submit application

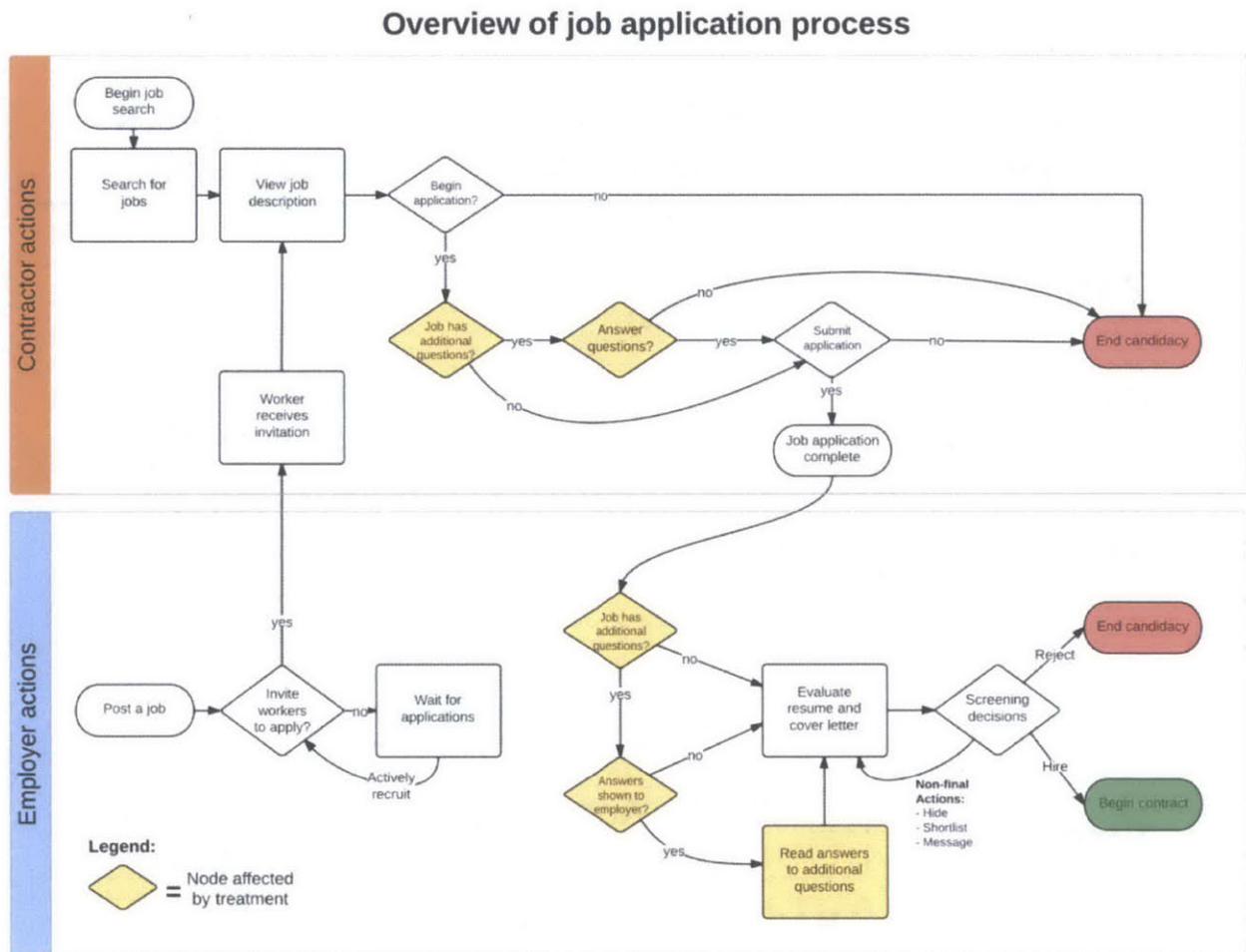
Notes: Each job application requires that workers submit a cover letter and propose a wage. Workers who are assigned to the treatment cells with questions are also required to answer those questions.

Figure 1-3: Experimental treatments



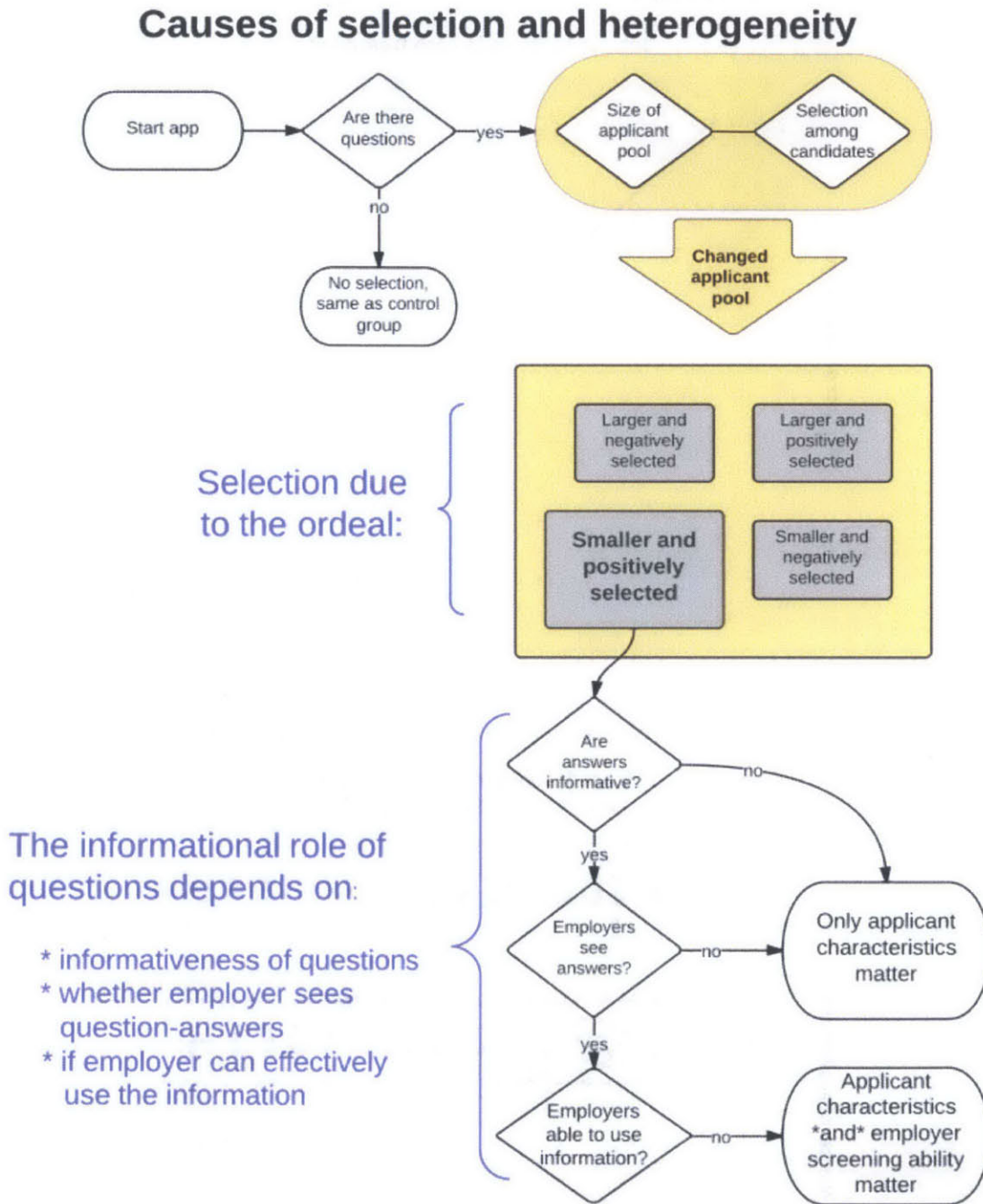
Notes: In the above matrix, the rows indicate the number of questions required and the columns indicate whether the question-answers were shown to employers. The most important comparisons are between the control group, Q5-NoAns, and Q5-Ans. The *information effect* is identified by comparing Q5-Ans and Q5-NoAns. In these treatments, the ordeal is held constant while information varies. The *selection effect* is identified by comparing control with Q5-NoAns. In these treatments, the amount of information provided to employers is held constant (at zero) while Q5-Ans imposes an ordeal.

Figure 1-4: Flowchart of application process and the effect of the experiment



Notes: The above provides an overview of the job application process. Diamonds indicate decision points, rectangles represent activities, and ovals indicate initial and terminal nodes. Parts of the process that are affected by the randomization are indicated in yellow.

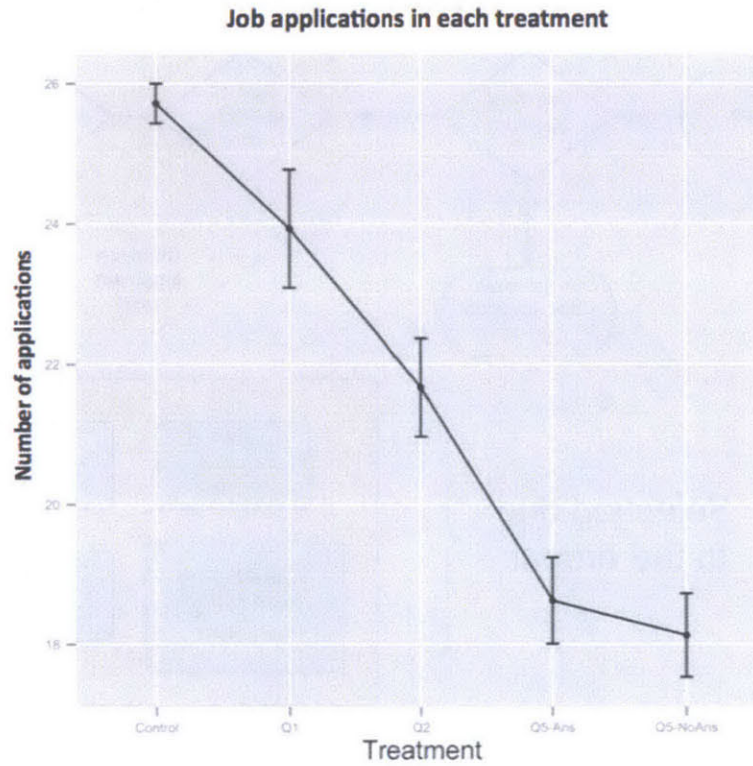
Figure 1-5: Flowchart of selection process and the effect of the experiment



Notes: The above provides an overview of the causal mechanisms at work during the experiment. The blocks of blue text describe the causal channels in terms of selection and information. Diamonds indicate decision points, rectangles represent activities, and ovals indicate initial and terminal nodes. Parts of the process that are affected by the randomization are indicated in yellow.

Figure 1-6: Effect of treatment on number of applications

(a) Overall effect of treatment on job applications



(b) Heterogeneous effect of treatment on job applications

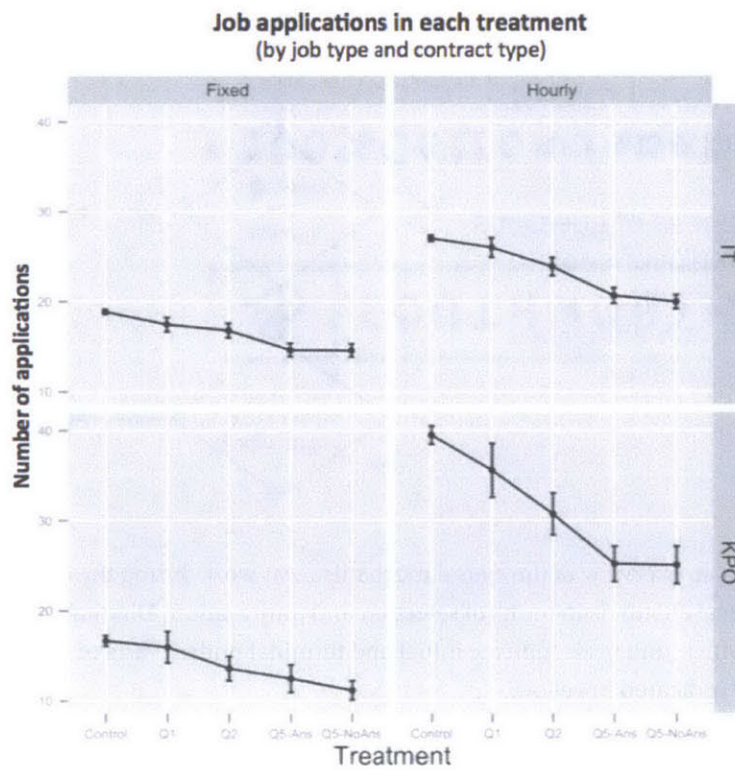


Table 1-1: List of questions used in treatment

Type of question	Specific question
Fit	Why do you think you are a good fit for this particular project? Why did you apply to this particular job? What part of this project most appeals to you? What questions do you have about the project?
Skill/expertise	What past project or job have you had that is most like this one and why? Which of the required job skills do you feel you are strongest at? What challenging part of this job are you most experienced in? Have you taken any oDesk tests and done well on them that you think are relevant to this job?
Project-specific knowledge	Do you have suggestions to make this project run successfully? Which part of this project do you think will take the most time?

Notes: This is the full set of questions that were randomly chosen and shown to applicants. For vacancies assigned to receive questions, all candidates to the vacancy were required to provide answers to the same set of questions.

Table 1-2: Summary statistics for workers

Variable	Mean	SD	25th	50th	75th	N
Male	0.692					169,987
Age	30.5	8.1	25	29	34	37,378
North America	0.132					222,671
East Asia	0.228					222,671
South Asia	0.441					222,671
Eastern Europe	0.042					222,671
Western Europe	0.076					222,671
Australasia	0.009					222,671
College educated	0.753					173,469
Experience (yrs)	5.92	4.58	3	5	7	153,138
>5 Years Experience (%)	0.56					153,138
Days on oDesk	368	448	14	191	578	222,671
New to oDesk	0.253					222,671
Profile Wage*	10.59	11.98	3	7	13.5	213,915
Avg Hourly Wage	7.92	8.13	2.22	5.15	10.79	69,127
Avg size of FP job	78.38	185.95	1.75	18	65.2	77,659
Any hires	0.435					222,671
Total hires (if > 0)	12.5	18.7	2	5	15	96,850
Any past earnings	0.388					222,671
Hours worked (if > 0)	657	1176	25	155	707	69,127
Total Earnings (if > 0)	4537	10399	80	620	3543	86,350
Earnings: Hourly (if > 0)	4854	10950	104	715	3925	69,969
Earnings: FP (if > 0)	853	1929	28	153	698	62,013
Average FB score	4.59	0.72	4.5	4.89	5	70,756

* Self-reported

Notes: This table describes characteristics for the 222,671 workers who applied or were invited to a job posted by employers in our sample. All characteristics are measured at the time the worker was first invited or applied. *New to oDesk* indicates that the contracted joined in the past two weeks. The *profile wage* (desired hourly wage) and *years of experience* are self-reported by workers in their public profiles. *Avg hourly wage* and *Avg FP earnings* represent the mean hourly wage and average earnings for each FP contract (for those who have worked). *Any hires* and *Any past earnings* indicate whether a worker has been hired or earned wages. Since so many workers are new to oDesk, we report conditional-on-positive summary statistics for earnings, hires, and hours worked. Finally, *Average FB score* indicates mean feedback on a scale of 1 to 5.

Table 1-3: Summary statistics for employers

Variable	Mean	SD	25th	50th	75th	N
Male	0.789					90,336
North America	0.575					89,940
Western Europe	0.158					89,940
Australasia	0.098					89,940
East Asia	0.042					89,940
South Asia	0.047					89,940
Days on oDesk	126	202	0	1	202	89,940
New to oDesk	0.585					89,940
Any past posts	0.401					90,336
# Past posts (if > 0)	10.8	19.6	2	4	12	36,228
Any hires	0.306					90,336
Number of hires (if > 0)	11	19.6	2	5	12	27,674
Any spend	0.295					90,336
Total spend (if > 0)	2,764	7,226	140	548	2,036	26,651

Notes: This table describes characteristics for the 90,336 employers in our sample at the time of their first job post. Geographic locations and the date a user joined is missing for 396 people, leaving only 89,940 observations for those variables. *New to oDesk* indicates that the employer joined in the past two weeks. The *profile wage* is the desired hourly wage stated by workers in their profile. Since so many employers are new to oDesk, we report conditional-on-positive summary statistics for the number of past job posts, past hires, and total spending.

Table I-4: Covariate balance: Employer characteristics by treatment cell

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	New to oDesk	Any Spend	Log total spend	Any past hires	Total # of hires	Has past posts	# of past posts	North American
Q1-Ans	0.003 (0.007)	-0.003 (0.006)	-0.067 (0.049)	-0.005 (0.006)	-0.165 (0.200)	-0.004 (0.006)	-0.118 (0.183)	-0.003 (0.007)
Q2-Ans	0.001 (0.006)	-0.000 (0.006)	-0.065 (0.047)	-0.001 (0.006)	-0.292* (0.143)	0.001 (0.006)	-0.151 (0.166)	0.007 (0.006)
Q5-Ans	0.004 (0.007)	-0.002 (0.006)	0.085 (0.047)	-0.000 (0.006)	-0.108 (0.159)	-0.004 (0.006)	0.140 (0.198)	0.002 (0.007)
Q5-NoAns	0.007 (0.006)	-0.005 (0.006)	0.031 (0.048)	-0.003 (0.006)	0.090 (0.177)	-0.004 (0.006)	0.129 (0.180)	-0.007 (0.006)
Control	0.584*** (0.002)	0.296*** (0.002)	6.260*** (0.014)	0.307*** (0.002)	3.492*** (0.062)	0.402*** (0.002)	4.380*** (0.061)	0.576*** (0.002)
F-test (p-value)	0.797	0.918	0.087	0.903	0.248	0.898	0.633	0.560
N	89,940	90,336	26,651	90,336	90,336	90,336	90,336	89,940
R ²	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table examines whether employer characteristics were balanced across our treatments by showing OLS regressions of each variable on treatment dummies. A description of the meaning of variables can be found in table I-3. Since our randomization occurred at the level of the employer and each observation is an employer, we do not cluster standard errors.

Table 1-5: Applicant pool selection by treatment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Worker quality index	Log days on oDesk	Log profile wage	Any past hires	Log # of past hires	Log hours worked	Log total earnings	Std. oDesk FB score	North American
One question	0.000177* (0.0000706)	0.00289 (0.0103)	0.0182 (0.0176)	0.00471 (0.00352)	0.0207* (0.00926)	0.00519 (0.0170)	0.0263 (0.0275)	0.00627 (0.00359)	0.00165 (0.00147)
Two questions	0.000281*** (0.0000830)	0.0253* (0.0106)	0.0606*** (0.0175)	0.0123*** (0.00344)	0.0358*** (0.00905)	0.0180 (0.0177)	0.0894*** (0.0269)	0.00148 (0.00374)	0.00481** (0.00168)
Five questions (includes Q5-Ans and Q5-NoAns)	0.000246*** (0.0000552)	0.0407*** (0.00768)	0.0812*** (0.0127)	0.0182*** (0.00240)	0.0601*** (0.00662)	0.0250 (0.0132)	0.125*** (0.0197)	0.00105 (0.00293)	0.00472*** (0.00122)
Control	0.0163*** (0.0000212)	5.522*** (0.00301)	1.888*** (0.00541)	0.783*** (0.000985)	2.200*** (0.00269)	5.067*** (0.00508)	6.647*** (0.00788)	-0.000868 (0.00106)	0.0547*** (0.000446)
Tests (and p-values) for:									
Delta Q5 (Q5-Ans minus Q5-NoAns)	.659	.5	.276	.56	.529	.675	.384	.339	.952
Ordeal size	.571	.008	.008	.003	.001	.62	.009	.457	.183
N	1,849,933	1,823,918	1,818,841	1,849,933	1,453,852	1,187,205	1,368,484	1,197,714	1,849,933
Clusters	74,463	74,378	74,404	74,463	73,511	71,920	73,140	72,468	74,463
R ²	0.000135	0.0000877	0.000972	0.000224	0.000225	0.0000168	0.000299	0.0000255	0.0000614

Robust standard errors are clustered at the job opening level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table examines applicant pool selection by regressing worker characteristics on the number of questions in each treatment. Rather than report individual treatments, we refer the treatments by the number of questions: "One question" = Q1-Ans, "Two questions" = Q2-Ans, "Five questions" = Q5-Ans and Q5-NoAns. Finally, the row "Ordeal size" reports p-values from a test of equality of the one-, two-, and five-question treatment indicators, which determines whether ordeal size caused differential selection. Additionally, the row "Delta Q5" reports a test of equality for the Q5-Ans and Q5-NoAns treatments; since these treatments appeared identical to applicants, there should fail to reject the null of no difference. Each observation is a job application to openings that received applications and we cluster standard errors are clustered at the job opening level.

Table 1-6: Hiring outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Number of applications	Log applications	Any invitations	Number of Invitations	"Fill rate" (any hires)	Hire from applicant pool	Worker quality index	Log profile wage of hire	Log budget of FP contract	Log hourly wage of hired
Q1-Ans	-1.746*** (0.457)	-0.068*** (0.015)	0.004 (0.007)	-0.170* (0.083)	-0.006 (0.006)	-0.005 (0.006)	.000109 (.000275)	-0.028 (0.018)	-0.071 (0.046)	-0.028 (0.029)
Q2-Ans	-4.071*** (0.385)	-0.145*** (0.015)	0.008 (0.006)	-0.132 (0.081)	-0.007 (0.006)	-0.004 (0.006)	.000472 (.000275)	-0.006 (0.018)	0.028 (0.045)	0.032 (0.029)
Q5-Ans	-7.008*** (0.346)	-0.280*** (0.015)	0.001 (0.007)	0.110 (0.144)	-0.015* (0.006)	-0.011 (0.006)	-.000240 (.000277)	-0.023 (0.018)	0.039 (0.048)	-0.037 (0.029)
Q5-NoAns	-7.475*** (0.338)	-0.298*** (0.015)	0.006 (0.006)	0.231 (0.181)	-0.014* (0.006)	-0.021*** (0.006)	.000402 (.000276)	0.014 (0.017)	0.041 (0.047)	0.007 (0.028)
Control	25.677*** (0.144)	2.711*** (0.004)	0.413*** (0.002)	2.039*** (0.038)	0.393*** (0.002)	0.318*** (0.002)	.021742*** (.000081)	2.116*** (0.005)	3.987*** (0.014)	1.980*** (0.008)
Information effect	0.467 (0.439)	0.018 (0.020)	-0.005 (0.009)	-0.121 (0.225)	-0.002 (0.009)	0.010 (0.008)	-.000643 (.000374)	-0.037 (0.024)	-0.002 (0.065)	-0.044 (0.038)
N	90,336	86,627	90,336	90,336	90,336	90,336	19,806	34,785	16,362	18,830
R ²	0.006	0.008	0.000	0.000	0.000	0.000	.000	0.000	0.000	0.000

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table reports a variety of outcomes related to hiring. Each column is a particular outcome and is regressed on all of the different treatments. The row "Information effect" reports the difference of the Q5-Ans and Q5-NoAns treatments and represents the effect of providing information to employers. *Any invitations* indicates whether the employer sent one or more invitations, "Fill rate" indicates whether the employer made a hire for the vacancy, and "Hire from applicant pool" indicates whether the person hired was a *non-invited* candidate who came from the ordinary applicant pool. Each observation represents the first job post by an employer and, since our randomization occurred at the level of the employer, we do not cluster standard errors.

Table 1-7: Employment outcomes: Job performance

	(1)	(2)	(3)	(4)	(5)
	Contract ended successfully	Std. oDesk FB score	Employer paid bonus	FP contract over budget	Employer asked for refund
Q1-Ans	0.017 (0.011)	-0.033 (0.035)	-0.001 (0.004)	0.017 (0.015)	-0.004 (0.005)
Q2-Ans	0.011 (0.011)	-0.019 (0.033)	-0.001 (0.004)	0.025 (0.015)	0.002 (0.005)
Q5-Ans	0.028* (0.011)	0.010 (0.034)	0.004 (0.004)	0.030 (0.016)	-0.005 (0.005)
Q5-NoAns	-0.006 (0.011)	0.005 (0.035)	0.006 (0.004)	0.014 (0.015)	0.001 (0.005)
Control	0.506*** (0.003)	0.003*** (0.010)	0.050*** (0.001)	0.380*** (0.004)	0.064*** (0.002)
Information effect	0.034* (0.015)	0.0055 (0.047)	-0.002 (0.005)	0.015 (0.021)	-0.006 (0.007)
N	29,543	21,416	48,767	16,362	35,192
R ²	0.000	0.000	0.000	0.000	0.000

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table reports several measures of success for contracts that were formed from openings in our experiment. Columns 1 and 2 are oDesk measures and columns 3-5 are measures that we constructed. "Std. oDesk FB score" is the standardized 5-star feedback rating using the mean and standard deviation from the control group. "Contract ended successfully" is an indicator for whether employers ended a contract and reported a successful outcome. Column 3 reports whether a bonus was granted in an hourly contract and column 4 reports whether a FP contract was over-budget. Both of these measures may indicate that the employer was satisfied with work and expanded the scope of the project. Column 5 indicates whether employers were unhappy with a contract and requested a refund. As described in table 1-6, we do not cluster standard errors.

Table 1-8: Employment outcomes: Contract spend

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Any spend	Total spend	Log spend	Log(1+x) spend	Above median spend	Log spend at 90th quantile	QMLE- Poisson Net spend
Q1-Ans	-0.006 (0.006)	-6.397 (11.262)	-0.056 (0.038)	-0.038 (0.031)	-0.006 (0.005)	-0.070 (0.079)	-0.047 (0.085)
Q2-Ans	-0.003 (0.006)	6.450 (12.165)	0.014 (0.037)	-0.008 (0.031)	0.002 (0.005)	-0.003 (0.077)	0.045 (0.084)
Q5-Ans	-0.009 (0.006)	-2.991 (10.701)	0.061 (0.038)	-0.010 (0.031)	-0.005 (0.005)	0.137 (0.079)	-0.022 (0.079)
Q5-NoAns	-0.010 (0.006)	-7.939 (9.434)	0.003 (0.038)	-0.050 (0.030)	-0.005 (0.005)	0.048 (0.078)	-0.059 (0.071)
Control	0.345*** (0.002)	138.713*** (3.876)	4.576*** (0.011)	1.519*** (0.009)	0.195*** (0.002)	6.742*** (0.023)	4.932*** (0.028)
Information effect	0.001 (0.008)	4.949 (13.171)	0.058 (0.051)	0.040 (0.042)	0.001 (0.007)	0.090 (0.110)	0.037 (0.099)
N	90,336	90,336	29,617	90,336	90,336	29,617	90,336
R ²	0.000	0.000	0.000	0.000	0.000		

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This table reports the amount of money spent on contracts associated with each opening. Although an imperfect proxy, we use the amount spend on a contract as a measure of success. Columns 1-4 are self-explanatory. Column 5 is an indicator for whether a contract had above median spend where the median is defined for each job category and contract type. Column 6 reports a simple quantile regression of log spend measured at the 90th quantile and Column 7 reports a QMLE Poisson estimator for spend using robust standard errors (i.e., it is robust even when the mean is not equal to the variance). As described in table 1-6, we do not cluster standard errors.

Table 1-9: Heterogeneity of ordeal and information effects: by job category, contract type, and employer experience

	“Fill rate” (any hires)		Contract ended successfully		Log total spend	
	Ordeal	Info	Ordeal	Info	Ordeal	Info
	(1)	(2)	(3)	(4)	(5)	(6)
Entire sample	-0.014** (0.006)	-0.002 (0.006)	-0.006 (0.011)	0.034** (0.015)	0.003 (0.038)	0.058 (0.051)
Job category						
IT	-0.008 (0.008)	-0.007 (0.011)	0.002 (0.014)	0.024 (0.02)	-0.041 (0.048)	0.102 (0.065)
KPO	-0.023** (0.011)	0.008 (0.014)	-0.019 (0.018)	0.051** (0.024)	0.068 (0.061)	-0.003 (0.082)
Contract type						
Fixed	-0.026*** (0.009)	0.003 (0.013)	0.007 (0.016)	0.021 (0.022)	0 (0.049)	0.058 (0.067)
Hourly	-0.003 (0.009)	-0.006 (0.012)	-0.012 (0.015)	0.045** (0.02)	0.005 (0.052)	0.031 (0.07)
Employer experience						
High experience (above median hires)	-0.025 (0.018)	-0.018 (0.024)	-0.003 (0.025)	0.017 (0.034)	0.065 (0.084)	0.128 (0.119)
Low experience (below median hires)	-0.023 (0.016)	0.016 (0.022)	0.027 (0.024)	0.016 (0.032)	-0.083 (0.079)	0.072 (0.105)
No experience (no hires)	-0.008 (0.007)	-0.004 (0.01)	-0.019 (0.015)	0.046** (0.02)	0.012 (0.05)	0.029 (0.067)

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports the ordeal effect and heterogeneity effect by various job categories, contract types, and employer experiences. The odd columns show the ordeal effects and the even columns show the information effects. As described in table 1-6, we do not cluster standard errors.

Table 1-10: Applicant pool selection by treatment (additional variables)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Log average hourly wage	Log average FP contract size	Male	Log age	College Graduate	Current student	Over 5 years work experience
One question	0.0225 (0.0230)	0.00570 (0.0232)	-0.000221 (0.00299)	0.00125 (0.00153)	-0.00177 (0.00232)	0.000248 (0.00125)	0.00852* (0.00365)
Two questions	0.0691** (0.0221)	0.0623** (0.0228)	0.000771 (0.00309)	0.00423* (0.00170)	0.00308 (0.00227)	-0.00303* (0.00120)	0.0210*** (0.00374)
Five questions	0.0835*** (0.0164)	0.0917*** (0.0167)	0.00660** (0.00211)	0.000994 (0.00125)	0.00608*** (0.00172)	-0.00294** (0.000946)	0.0239*** (0.00266)
Control	1.544*** (0.00704)	3.447*** (0.00684)	0.744*** (0.000883)	3.362*** (0.000461)	0.792*** (0.000684)	0.0653*** (0.000375)	0.618*** (0.00113)
Tests (and p-values) for:							
Delta Q5 (<i>Q5-Ans</i> <i>minus Q5-NoAns</i>)	.265	.116	.816	.618	.125	.757	.628
Ordeal size	.069	.006	.076	.245	.015	.065	.001
N	1,181,267	1,044,605	1,483,572	581,143	1,535,206	1,535,206	1,584,856
Clusters	71,888	72,102	73,768	67,102	73,740	73,740	73,822
R ²	0.000706	0.000344	0.0000223	0.0000223	0.0000259	0.0000211	0.000321

Robust clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table shows applicant pool selection by *number of questions*. It has an identical format to table 1-5, except that it shows additional, less-important variables. Rather than list results by treatment, we increase power by grouping together the two five-question treatments since they appear identical to workers. The row “Delta Q5” tests for equality between the two five-question treatments and the row “Ordeal size” tests for equality between the one-, two-, and five-question treatments in order to measure whether more questions led to different amounts of selection. Each observation is a job application to a publicly listed opening from experiment. Table 1-2 describes each variable. Finally, we cluster standard errors at the level of job opening for reasons described in table 1-2.

Table 1-11: Comparison of information effect controlling for worker quality index

(a) Performance outcomes

	(1)	(2)	(3)	(4)	(5)
	Contract ended successful	Std. oDesk FB score	Employer paid bonus	FP contract over budget	Employer asked for refund
Information effect					
Unadjusted	0.032 (0.020)	-0.004 (0.047)	-0.010 (0.016)	0.001 (0.025)	0.001 (0.008)
Controlling for quality index	0.034 (0.020)	0.006 (0.047)	-0.009 (0.016)	0.002 (0.025)	0.000 (0.008)
<i>N</i>	15,996	11,378	10,065	9,741	19,806

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(b) Spend outcomes

	(1)	(2)	(3)	(4)	(5)
	Any billing	Total billings	Log billings	Log(1+x) billings	Above median billings
Information effect					
Unadjusted	0.008 (0.012)	21.106 (41.547)	0.087 (0.062)	0.129 (0.079)	-0.000 (0.018)
Controlling for quality index	0.009 (0.012)	19.202 (41.536)	0.086 (0.062)	0.133 (0.079)	0.002 (0.018)
<i>N</i>	19,806	19,806	16,429	19,806	19,806

Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table reports the information effects for various outcomes along with estimates that adjust for the worker quality index of each hire. By controlling for the worker quality index of the worker hired, we are able to isolate the effect of information by controlling for the initial quality of workers. The worker quality index is as defined in Section 1.3.2 and all outcomes are defined as in Table 1-7 and Table 1-8. The unadjusted information effect estimates are slightly different from those in tables 1-7 and 1-8 since, in constructing our quality-index, we are forced to use a slightly different sample.

Chapter 2

Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets

Abstract

We conduct the first natural field experiment to explore the relationship between the “meaningfulness” of a task and worker effort. We employed about 2,500 workers from Amazon’s Mechanical Turk (MTurk), an online labor market, to label medical images. Although given an identical task, we experimentally manipulated how the task was framed. Subjects in the *meaningful* treatment were told that they were labeling tumor cells in order to assist medical researchers, subjects in the *zero-context* condition (the control group) were not told the purpose of the task, and, in stark contrast, subjects in the *shredded* treatment were not given context and were additionally told that their work would be discarded. We found that when a task was framed more meaningfully, workers were more likely to participate. We also found that the meaningful treatment increased the quantity of output (with an insignificant change in quality) while the shredded treatment decreased the quality of output (with no change in quantity). We believe these results will generalize to other short-term labor markets. Our study also discusses MTurk as an exciting platform for running natural field experiments in economics.

2.1 Introduction

Economists, philosophers, and social scientists have long recognized that non-pecuniary factors are powerful motivators that influence choice of occupation. For a multidisciplinary literature review on the role of meaning in the workplace, we recommend Rosso et al. (2010). Previous studies in this area have generally been based on ethnographies, observational studies, or laboratory experiments. For instance, Wrzesniewski et al. (1997) used ethnographies to classify work into jobs, careers, or callings. Using an observation study, Preston (1989) demonstrated

that workers may accept lower wages in the non-profit sector in order to produce goods with social externalities. Finally, Ariely et al. (2008) showed that labor had to be both recognizable and purposeful to have meaning. In this paper, we limit our discussion to the role of meaning in economics, particularly through the lens of competing differentials. We perform the first *natural field experiment* (Harrison and List, 2004) in a real effort task that manipulates levels of meaningfulness. This method overcomes a number of shortcomings of the previous literature, including: interview bias, omitted variable bias, and concerns of external validity beyond the laboratory.

We study whether employers can deliberately alter the perceived “meaningfulness” of a task in order to induce people to do more and higher quality work and thereby work for a lower wage. We chose a task that would appear meaningful for many people if given the right context — helping cancer researchers mark tumor cells in medical images. Subjects in the *meaningful* treatment were told the purpose of their task is to “help researchers identify tumor cells;” subjects in our *zero-context* group were not given any reason for their work and the cells were instead referred to as mere “objects of interest” and laborers in the *shredded* group were given zero context but also explicitly told that their labelings would be discarded upon submission. Hence, the pay structure, task requirements, and working conditions were identical, but we added cues to alter the perceived meaningfulness of the task.

We recruited workers from the United States and India from Amazon’s Mechanical Turk (MTurk), an online labor market where people around the world complete short, “one-off” tasks for pay. The MTurk environment is a spot market for labor characterized by relative anonymity and a lack of strong reputational mechanisms. As a result, it is well-suited for an experiment involving the meaningfulness of a task since the variation we introduce regarding a task’s meaningfulness is less affected by desires to exhibit pro-social behavior or an anticipation of future work (career concerns). We ensured that our task appeared like any other task in the marketplace and was comparable in terms of difficulty, duration, and wage.

Our study is representative of the kinds of natural field experiments for which MTurk is particularly suited. Section 2.2.2 explores MTurk’s potential as a platform for field experimentation using the framework proposed in Levitt and List (2009, 2007).

We contribute to the literature on compensating wage differentials (Rosen, 1986) and the organizational behavioral literature on the role of meaning in the workplace (Rosso et al., 2010). Within economics, Stern (2004) provides quasi-experimental evidence on compensating differentials within the labor market for scientists by comparing wages for academic and private sector job offers among recent Ph.D. graduates. He finds that “scientists pay to be scientists”

and require higher wages in order to accept private sector research jobs because of the reduced intellectual freedom and a reduced ability to interact with the scientific community and receive social recognition. Ariely et al. (2008) use a laboratory experiment with undergraduates to vary the meaningfulness of two separate tasks: (1) assembling Legos and (2) finding 10 instances of consecutive letters from a sheet of random letters. Our experiment augments experiment 1 in Ariely et al. (2008) by testing whether their results extend to the field. Additionally, we introduce a richer measure of task effort, namely *task quality*. Where our experiments are comparable, we find that our results parallel theirs.

We find that the main effects of making our task more meaningful is to induce a higher fraction of workers to complete our task, hereafter dubbed as “induced to work.” In the meaningful treatment, 80.6% of people labeled at least one image compared with 76.2% in the zero-context and 72.3% in the shredded treatments.

After labeling their first image, workers were given the opportunity to label additional images at a declining piecerate. We also measure whether the treatments increase the quantity of images labeled. We classify participants as “high-output” workers if they label five or more images (an amount corresponding to roughly the top tercile of those who label) and we find that workers are approximately 23% more likely to be high-output workers in the meaningful group.

We introduce a measure of task quality by telling workers the importance of accurately labeling each cell by clicking as close to the center as possible. We first note that MTurk labor is high quality, with an average of 91% of cells found. The meaning treatment had an ambiguous effect, but the shredded condition in both countries lowered the proportion of cells found by about 7%.

By measuring both quantity and quality we are able to observe how task effort is apportioned between these two “dimensions of effort.” Do workers work “harder” or “longer” or both? We found an interesting result: the meaningful condition seems to increase quantity without a corresponding increase in quality and the shredded treatment decreases quality without a corresponding decrease in quantity. Investigating whether this pattern generalizes to other domains may be a fruitful future research avenue.

Finally, we calculate participants’ average hourly wage based on how long they spent on the task. We find that subjects in the meaningful group work for \$1.34 per hour, which is 6 cents less per hour than zero context participants and 14 cents less per hour than shredded condition participants.

We expect our findings to generalize to other short-term work environments such as temporary employment or piecework. In these environments, employers may not consider that

non-pecuniary incentives of meaningfulness matter; we argue that these incentives do matter, and to a significant degree.

Section 2.2 provides background on MTurk and discusses its use as a platform for conducting economic field experiments. Section 2.3 describes our experimental design. Section 2.4 presents our results and discussion and Section 2.5 concludes. Appendices A provides full details on our experimental design.

2.2 Mechanical Turk and its potential for field experimentation

Amazon's Mechanical Turk (MTurk) is the largest online, task-based labor market and is used by hundreds of thousands of people worldwide. Individuals and companies can post tasks (known as Human Intelligence Tasks, or "HITs") and have them completed by an on-demand labor force. Typical tasks include image labeling, audio transcription, and basic internet research. Academics also use MTurk to outsource low-skilled resource tasks such as identifying linguistic patterns in text (Sprouse, 2011) and labeling medical images (Holmes and Kapelner, 2010). The image labeling system from the latter study, known as "DistributeEyes", was originally used by breast cancer researchers and was modified for our experiment.

Beyond simply using MTurk as a source of labor, academics have also begun using MTurk as a way to conduct online experiments. The remainder of the section highlights some of the ways this subject pool is used and places special emphasis on the suitability of the environment for natural field experiments in economics.

2.2.1 General use by social scientists

As Henrich et al. (2010) argue, many findings from social science are disproportionately based on what he calls "W.E.I.R.D." subject pools (**W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic) and as a result it is inappropriate to believe the results generalize to larger populations. Since MTurk has users from around the world, it is also possible to conduct research across cultures. For example, Eriksson and Simpson (2010) use a cross-national sample from MTurk to test whether differential preferences for competitive environments are explained by females' stronger emotional reaction to losing, hypothesized by Croson and Gneezy (2009).

It is natural to ask whether results from MTurk generalize to other populations. Paolacci et al. (2010) assuage these concerns by replicating three classic framing experiments on MTurk: The Asian Disease Problem, the Linda Problem and the Physician Problem; Horton et al. (2011)

provide additional replication evidence for experiments related to framing, social preferences, and priming. Berinsky et al. (2012) argues that the MTurk population has “attractive characteristics” because it approximates gold-standard probability samples of the US population. All three studies find that the direction and magnitude of the effects line up well compared with those found in the laboratory.

An advantage of MTurk relative to the laboratory is that the researcher can rapidly scale experiments and recruit hundreds of subjects within only a few days and at substantially lower costs.¹

2.2.2 Suitability for natural field experiments in Economics

Apart from general usage by academics, the MTurk environment offers additional benefits for experimental economists and researchers conducting natural field experiments. We analyze the MTurk environment within the framework laid out in Levitt and List (2009, 2007).

In the ideal natural field experiment, “the environment is such that the subjects naturally undertake these tasks and [do not know] that they are participants in an experiment.” Additionally, the experimenter must exert a high degree of control over the environment without attracting attention or causing participants to behave unnaturally. MTurk’s power comes from the ability to construct customized and highly-tailored environments related to the question being studied. It is possible to collect very detailed measures of user behavior such as precise time spent on a webpage, mouse movements, and positions of clicks. In our experiment, we use such data to construct a precise quality measure.

MTurk is particularly well-suited to using experimenter-as-employer designs (Gneezy and List, 2006) as a way to study worker incentives and the employment relationship without having to rely on cooperation of private sector firms.² For example, Barankay (2010) posted identical image labeling tasks and varied whether workers were given feedback on their relative performance (i.e., ranking) in order to study whether providing rank-order feedback led workers to return for a subsequent work opportunity. For a more detailed overview of how online labor markets can be used in experiments, see Horton et al. (2011).

Levitt and List (2007) enumerate possible complications that arise when experimental findings are extrapolated outside the lab: *scrutiny*, *anonymity*, *stakes*, *selection*, and *artificial re-*

¹For example, in our study we paid 2,471 subjects \$789 total and they worked 701 hours (equating to 31 cents per observation). This includes 60 subjects whose data were not usable.

²Barankay (2010) remarks that “the experimenter [posing] as the firm [gives] substantial control about the protocol and thereby eliminates many project risks related to field experiments.

strictions. We analyze each complication in the context of our experiment and in the context of experimentation using MTurk in general.

Scrutiny and anonymity. In the lab, experimenter effects can be powerful; subjects behave differently if they are aware their behavior is being watched. Relatedly, subjects frequently lack anonymity and believe their choices will be scrutinized after the experiment. In MTurk, interaction between workers and employers is almost non-existent; most tasks are completed without any communication and workers are only identifiable by a numeric identifier. Consequently, we believe that MTurk experiments are less likely to be biased by these complications.

Stakes. In the lab or field, it's essential to "account properly for the differences in stakes across settings" (Levitt and List, 2007). We believe that our results would generalize to other short-term work environments, but would not expect them to be generalizable to long-term employment decisions such as occupational choice. Stakes must also be chosen adequately for the environment and so we were careful to match wages to the market average.

Selection. Experiments fail to be generalizable when "participants in the study differ in systematic ways from the actors engaged in the targeted real-world setting." We know that within MTurk, it is unlikely that there is selection into our experiment since our task was designed similar in appearance to real tasks. The MTurk population also seems representative along a number of observable demographic characteristics (Berinsky et al., 2012); however, we acknowledge that there are potentially unobservable differences between our subject pool and the broader population. Still, we believe that MTurk subject behavior would generalize to workers' behavior in other short-term labor markets.

Artificial restrictions. Lab experiments place unusual and artificial restrictions on the actions available to subjects and they examine only small, non-representative windows of time because the experimenter typically doesn't have subjects and time horizons for an experiment. In structuring our experiment, workers had substantial latitude in how they performed their task. In contrast with the lab, subjects could "show-up" to our task whenever they wanted, leave at will, and were not time-constrained. Nevertheless, we acknowledge that while our experiment succeeded in matching short-term labor environments like MTurk, that our results do not easily generalize to longer-term employment relationships.

Levitt and List (2009) highlight two limitations of field experiments vis-a-vis laboratory experiments: the *need for cooperation* with third parties and the difficulty of *replication*. MTurk does not suffer from these limitations. Work environments can be created by researchers without the need of a private sector partner, whose interests may diverge substantially from that of the researcher. Further, MTurk experiments can be replicated simply by downloading source

code and re-running the experiment. In many ways, this allows a push-button replication that is far better than that offered in the lab.

2.3 Experimental Design

2.3.1 Subject recruitment

In running our randomized natural field experiment, we posted our experimental task so that it would appear like any other task (image labeling tasks are among the most commonly performed tasks on MTurk). Subjects had no indication they were participating in an experiment. Moreover, since MTurk is a market where people ordinarily perform one-off tasks, our experiment could be listed inconspicuously.

We hired a total of 2,471 workers (1,318 from the US and 1,153 from India). Although we tried to recruit equally from both countries, there were fewer Indians in our sample since attrition in India was higher. We collected each worker's age and gender during a "colorblindness" test that we administered as part of the task. These and other summary statistics can be found in Table 2-1. By contracting workers from the US and India, we can also test whether workers from each country respond differentially to the meaningfulness of a task.

Our task was presented so that it appeared like a one-time work opportunity (subjects were barred from doing the experiment more than once) and our design sought to maximize the amount of work we could extract during this short interaction. The first image labeling paid \$0.10, the next paid \$0.09, etc, leveling off at \$0.02 per image. This wage structure was also used in Ariely et al. (2008) and has the benefit of preventing people from working too long.

2.3.2 Description of experimental conditions

Upon accepting our task, workers provided basic demographic information and passed a color-blindness test. Next, they were randomized into either the *meaningful*, the *zero-context*, or the *shredded* condition. Those in the shredded condition were shown a warning message stating that their labeling will not be recorded and we gave them the option to leave. Then, all participants were forced to watch an instructional video which they could not fast-forward. See the online supplement for the full script of the video as well as screenshots.

The video for the meaningful treatment began immediately with cues of meaning. We adopt a similar working definition of "meaningfulness" as used in Ariely et al. (2008): "Labor [or a task] is meaningful to the extent that (a) it is recognized and/or (b) has some point or purpose."

We varied the levels of meaningfulness by altering the degree of recognition and the detail used to explain the purpose of our task. In our meaningful group, we provided “recognition” by thanking the laborers for working on our task. We then explained the “purpose” of the task by creating a narrative explaining how researchers were inundated with more medical images than they could possibly label and that they needed the help of ordinary people. In contrast, the zero-context and shredded groups were not given recognition, told the purpose of the task, or thanked for participating; they were only given basic instructions. Analyzing the results from a post-manipulation check (see section 2.4.4), we are confident that these cues of meaning induced the desired affect.

Both videos identically described the wage structure and the mechanics of how to label cells and properly use the task interface (including zooming in/out and deleting points, which are metrics we analyze). However, in the meaningful treatment, cells were referred to as “cancerous tumor cells” whereas in the zero-context and shredded treatments, they were referred to as nondescript “objects of interest.” Except for this phrase change, both scripts were identical during the instructional sections of the videos. To emphasize these cues, workers in the meaningful group heard the words “tumor,” “tumor cells,” “cells,” etc. 16 times before labeling their first image and similar cues on the task interface reminded them of the purpose of the task as they labeled.

2.3.3 Task interface, incentive structure, and response variables

After the video, we administered a short multiple-choice quiz testing workers’ comprehension of the task and user interface. In the shredded condition, we gave a final question asking workers to again acknowledge that their work will not be recorded.

Upon passing the quiz, workers were directed to a task interface which displayed the image to be labeled and allowed users to mark cancerous tumor cells (or “objects of interest”) by clicking (see figure 2-1). The image shown was one of ten look-alike photoshopped images displayed randomly. We also provide the workers with controls — *zoom functionality* and the ability to *delete points* — whose proper use would allow them to produce high-quality labelings.

During the experiment, we measured three response variables: (1) induced to work, (2) quantity of image labelings, and (3) quality of image labelings.

Many subjects can – and – do stop performing a task even after agreeing to complete it. While submitting bad work on MTurk is penalized, workers can abandon a task with only nominal penalty. Hence, we measure attrition with the response variable *induced to work*. Workers were only counted as induced to work if they watched the video, passed the quiz, and completed

one image labeling. Our experimental design deliberately encourages attrition by imposing an upfront and unpaid cost of watching a three-minute instructional video and passing a quiz before moving on to the actual task.

Workers were paid \$0.10 for the first image labeling. They were then given an option to label another image for \$0.09, and then another image for \$0.08, and so on.³ At \$0.02, we stopped decreasing the wage and the worker was allowed to label images at this pay rate indefinitely. After each image, the worker could either collect what they had earned thus far, or label more images. We used the *quantity of image labelings* for our second response variable.

In our instructional video, we emphasized the importance of marking the exact center of each cell. When a worker labeled a cell by clicking on the image, we measured that click location to the nearest pixel. Thus, we were able to detect if the click came “close” to the actual cell. Our third response variable, *quality of image labelings* is the proportion of objects identified based on whether a worker’s click fell within a pixel radius from the object’s true center. We will discuss the radii we picked in the following section.

After workers chose to stop labeling images and collect their earnings, they were given a five-question PMC survey which asked whether they thought the task (a) was enjoyable (b) had purpose (c) gave them a sense of accomplishment (d) was meaningful (e) made their efforts recognized. Responses were collected on a five-point Likert scale. We also provided a text box to elicit free-response comments.⁴

2.3.4 Hypotheses

Hypothesis 1 We hypothesize that at equal wages, the meaningful treatment will have the highest proportion of workers induced to work and the shredded condition will have the lowest proportion. In the following section, we provide theoretical justification for this prediction.

Hypothesis 2 As in Ariely et al. (2008), we hypothesize that *quantity* of images labeled will be increasing in the level of meaningfulness.

Hypothesis 3 In addition to quantity, we measure the *quality* of image labelings and hypothesize that this is increasing in the level of meaningfulness.

³Each image was randomly picked from a pool of ten look-alike images.

⁴About 24% of respondents left comments (no difference across treatments).

Hypothesis 4 Based upon prior survey research on MTurk populations, we hypothesize that *Indian workers are less responsive to meaning*. Ipeirotis (2010) finds that Indians are more likely to have MTurk as a primary source of income (27% vs. 14% in the US). Likewise, people in the US are nearly twice as likely to report doing tasks because they are fun (41% vs. 20%). Therefore, one might expect financial motivations to be more important for Indian workers.⁵

2.4 Experimental Results and Discussion

We ran the experiment on $N = 2,471$ subjects (1,318 from the United States and 1,153 from India). Table 2-1 shows summary statistics for our response variables (induced to work, number of images, and quality), demographic variables, and hourly wage.

Broadly speaking, as the level of meaning increases, subjects are more likely to participate and they label more images and with higher quality. Across all treatments, US workers participate more often, label more images, and mark points with greater accuracy. Table 2-2 uses a heatmap to illustrate our main effect sizes and their significance levels by treatment, country, and response variable. Each cell indicates the size of a treatment effect relative to the control (i.e., zero context condition). Statistically significant *positive* effects are indicated using green fill where darker green indicates higher levels of significance. Statistically significant *negative* effects are indicated using red fill where darker red indicates higher levels of significance. Black text without fill indicates effects that are marginally significant ($p < 0.10$). Light gray text indicates significance levels above 0.10.

Overall, we observe that the meaningful condition induces an increase in quantity without significantly increasing quality, and the shredded condition induces a quality decrease with quantity remaining constant. This “checkerboard effect” may indicate that meaning plays a role in moderating how workers trade quantity for quality i.e. how their energy is channeled in the task.

We now investigate each response variable individually.

2.4.1 Labor Participation Results: “Induced to work”

We investigate how treatment and country affects whether or not subjects chose to do our task. Unlike in a laboratory environment, our subjects were workers in a relatively anonymous labor market and were not paid a “show-up fee.” On MTurk, workers frequently start but do not finish

⁵Although Horton et al. (2011) find that workers of both types are strongly motivated by money.

tasks; attrition is therefore a practical concern for employers who hire from this market. In our experiment, on average, 25% of subjects began, but did not follow-through by completing one full labeling.

Even in this difficult environment, we were able to increase participation among workers by roughly 4.6% by framing the task as more meaningful (see columns 1 and 2 of table 2-3). The effect is robust to including various controls for age, gender, and time of day effects. As a subject in the meaningful treatment told us, “It’s always nice to have [HITs] that take some thought and mean something to complete. Thank you for bringing them to MTurk.” The shredded treatment discouraged workers and caused them to work 4.0% less often but the effect was less significant ($p = 0.057$ without controls and $p = 0.082$ with controls). Thus, hypothesis 1 seems to be correct.

Irrespective of treatment, subjects from India completed an image 18.5% less often ($p < 0.001$) than subjects from the US. We were interested in interactions between country and treatment, so we ran the separate induced-to-work regression results by country (unshown). We did not find significant effects within the individual countries because we were underpowered to detect this effect when the sample size was halved. We find no difference in the treatment effect for induced to work between India and the United States ($p = 0.97$). This is inconsistent with hypothesis 4 where we predicted Indian subjects to respond more strongly to pecuniary incentives.

It is also possible that the effects for induced to work were weak because subjects could have still attributed meaning to the zero context and shredded conditions, a problem that will affect our results for quantity and quality as well. This serves to bias our treatment effects downward suggesting that the true effect of meaning would be larger. For instance, one zero-context subject told us, “I assumed the ‘objects’ were cells so I guess that was kind of interesting.” Another subject in the zero-context treatment advised us, “you could put MTurkers to good use doing similar work with images, e.g. in dosimetry or pathology ... and it would free up medical professionals to do the heftier work.”

2.4.2 Quantity Results: Number of images labeled

Table 2-1 shows that the number of images increased with meaning. However, this result is conditional on being induced to work and is therefore contaminated with selection bias. We follow Angrist (2001) and handle selection by creating a dummy variable for “did two or more labelings” and a dummy for “did five or more labelings” and use them as responses (other cutoffs produced similar results).

We find mixed results regarding whether the the level of meaningfulness affects the quantity of output. Being assigned to the meaningful treatment group *did* have a positive effect, but assignment to the shredded treatment did not result in a corresponding decrease in output.

Analyzing the outcome “two or more labelings,” column 3 of table 2-3 shows that the meaningful treatment induced 4.7% more subjects to label two or more images ($p < 0.05$). The shredded treatment had no effect. Analyzing the outcome “five or more labelings” (column 5), which we denote as “high-output workers,”⁶ the meaningful treatment was highly significant and induced 8.5% more workers ($p < 0.001$ with and without controls), an increase of nearly 23 percent, and the shredded treatment again has no effect.

Hypothesis 2 (quantity increases with meaningfulness) seems to be correct only when comparing the meaningful treatment to the zero-context treatment. An ambiguous effect of the shredded treatment on quantity is also reported by Ariely et al. (2008).

We didn’t find differential effects between the United States and India. In an unshown regression, we found that Americans were 9.5% more likely to label five or more images ($p < 0.01$) and Indians were 8.4% more likely to label five or more ($p < 0.05$). These two effects were not found to be different ($p = 0.84$) which is inconsistent with hypothesis 4 that Indians are more motivated by pecuniary incentives than Americans.

Interestingly, we also observed a number of “target-earners” who stopped upon reaching exactly one dollar in earnings. A mass of 16 participants stopped at one dollar, while one participant stopped at \$1.02 and not one stopped at \$0.98, an effect also observed by Horton and Chilton (2010). The worker who labored longest spent 2 hours and 35 minutes and labeled 77 images.

2.4.3 Quality Results: Accuracy of labeling

Quality was measured by the fraction of cells labeled at a distance of five pixels (“coarse quality”) and two pixels (“fine quality”) from their true centers. In presenting our results (see table 3-6), we analyze the treatment effects using our fine quality measure. The coarse quality regression results were similar, but the fine quality had a much more dispersed distribution.⁷

Our main result is that fine quality was 7.2% lower in the shredded treatment, but there wasn’t a large corresponding increase in the meaningful treatment.⁸ This makes sense; if the

⁶Labeling five or more images corresponds to the top tercile of quantity among people who were induced to work.

⁷The inter-quartile range of coarse quality overall was [93.3%, 97.2%] whereas the IQR of fine quality was overall [54.7%, 80.0%].

⁸One caveat with our quality results is that we only observe quality for people who were induced to work and

workers knew their labelings weren't going to be checked, there is no incentive to mark points carefully. This result was not different across countries (regression unshown). The meaningful treatment has a marginally significant effect only in the United States, where fine quality increased by 3.9% ($p = 0.092$ without controls and $p = 0.044$ with controls), but there was no effect in India. Thus, hypothesis 3 (quality increases with meaningfulness) seems to be correct *only* when comparing the shredded to the zero context treatment which is surprising.

Although Indian workers were less accurate than United States workers and had 5.3% lower quality ($p < 0.001$ and robust to controls), United States and Indian workers did not respond differentially to the shredded treatment ($p = 0.53$). This again is inconsistent with hypothesis 4.

Experience matters. Once subjects had between 6 and 10 labelings under their belt, they were 1.8% less accurate ($p < 0.01$), and if they had done more than 10 labelings, they were 14% less accurate ($p < 0.001$). This result may reflect negative selection — subjects who labeled a very high number of images were probably working too fast or not carefully enough.⁹ Finally, we found that some of the ten images were substantial harder to label accurately than others (a partial F-test for equality of fixed effects results in $p < 0.001$).

2.4.4 Post Manipulation Check Results

In order to understand how our treatments affected the perceived meaningfulness of the task, we gave a post manipulation check to all subjects who completed at least one image and did not abandon the task before payment. This data should be interpreted cautiously given that subjects who completed the tasks and our survey are *not* representative of all subjects in our experiment.¹⁰

We found that those in the meaningful treatment rated significantly higher in the post manipulation check in both the United States and India. Using a five-point Likert scale, we asked workers to rate the perceived level of meaningfulness, purpose, enjoyment, accomplishment, and recognition. In the meaningful treatment, subjective ratings were higher in all categories but the self-rated level of meaningfulness and purpose were the highest. The level of meaning-

selected into our experiment (we have “attrition bias”). Attrition was 4% higher in the shredded treatment and we presume that the people who opted out of labeling images would have labeled them with far worse quality had they remained in the experiment.

⁹Anecdotally, subjects from the shredded condition who submitted comments regarding the task were less likely to have expressed concerns about their accuracy. One subject from the meaningful group remarked that “[his] mouse was too sensitive to click accurately, even all the way zoomed in,” but we found no such apologies or comments from people in the shredded group.

¹⁰Ideally, we would have collected this information immediately after introducing the treatment condition. However, doing so would have compromised the credibility of our natural field experiment.

fulness was 1.3 points higher in the US and 0.6 points higher in the India; the level of perceived purposefulness was 1.2 points higher in America and 0.5 points higher in India. In the United States, the level of accomplishment only increased by 0.8 and the level of enjoyment and recognition increased by 0.3 and 0.5 respectively with a marginal increase in India. As a US participant told us, “I felt it was a privilege to work on something so important and I would like to thank you for the opportunity.”

We conclude that the meaningful frames accomplished their goal. Remarkably, those in the shredded treatment in either country did not report significantly lower ratings on any of the items in the post manipulation check. Thus, the shredded treatment may not have had the desired effect.

2.5 Conclusion

Our experiment is the first that uses a natural field experiment in a real labor market to examine how a task’s meaningfulness influences labor supply.

Overall, we found that the greater the amount of meaning, the more likely a subject is to participate, the more output they produce, the higher quality output they produce, and the less compensation they require for their time. We also observe an interesting effect: high meaning increases *quantity* of output (with an insignificant increase in quality) and low meaning decreases *quality* of output (with no change in quantity). It is possible that the level of perceived meaning affects how workers substitute their efforts between task quantity and task quality. The effect sizes were found to be the same in the US and India.

Our finding has important implications for those who employ labor in any short-term capacity besides crowdsourcing, such as temp-work or piecework. As the world begins to outsource more of its work to anonymous pools of labor, it is vital to understand the dynamics of this labor market and the degree to which non-pecuniary incentives matter. This study demonstrates that they do matter, and they matter to a significant degree.

This study also serves as an example of what MTurk offers economists: an excellent platform for high internal validity natural field experiments while evading the external validity problems that may occur in laboratory environments.

Figure 2-1: Main task portal for a subject in the meaningful treatment

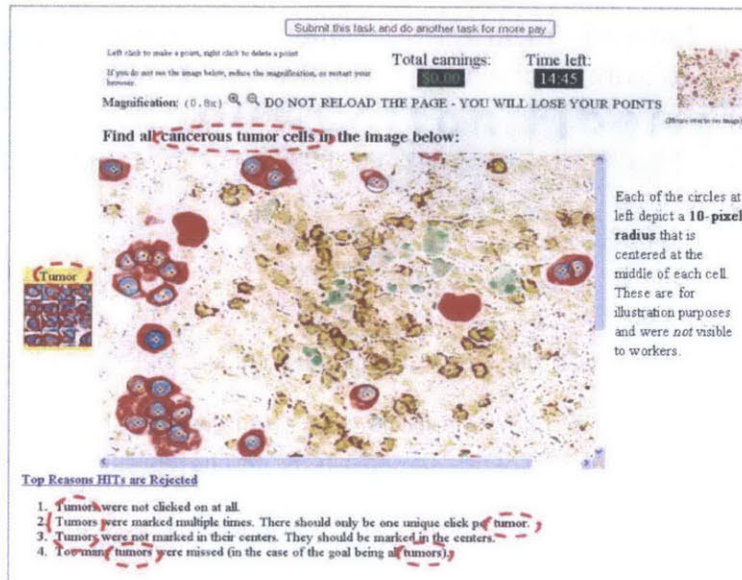


Table 2-1: Summary statistics for response variables and demographics by treatment and country

	Shredded	Zero Context	Meaningful	US only	India only
% Induced to Work	.723	.762	.806	.85	.666
# Images (if ≥ 1)	5.94 \pm 6.8	6.11 \pm 6.9	7.12 \pm 7.6	5.86 \pm 6.1	7.17 \pm 8.3
Did ≥ 2 labelings	.696	.706	.75	.797	.627
Did ≥ 5 labelings	.343	.371	.456	.406	.373
Avg Hourly Wage	\$1.49	\$1.41	\$1.34	\$1.50	\$1.29
% Male	.616	.615	.58	.483	.743
Age	29.6 \pm 9.3	29.6 \pm 9.5	29.3 \pm 9.1	31.8 \pm 10.5	26.9 \pm 6.8
N	828	798	845	1318	1153
Coarse quality	.883 \pm .21	.904 \pm .18	.930 \pm .14	.924 \pm .15	.881 \pm .21
Fine quality	.614 \pm .22	.651 \pm .21	.676 \pm .18	.668 \pm .19	.621 \pm .26
PMC Meaning	3.44 \pm 1.3	3.54 \pm 1.2	4.37 \pm 0.9	3.67 \pm 1.3	3.98 \pm 1.1

Notes: The statistics for the quality metrics are computed by averaging each worker's average quality (only for workers who labeled one or more images). The statistics for the PMC meaning question only include workers who finished the task and survey.

Table 2-2: A heatmap illustration of our results

	Induced to work	Did ≥ 5 labelings	Fine Quality	Average Hourly Wage
Meaningful	↑ 4.6%*	↑ 8.5%***	↑ 0.7%	↓ 4.5%
Meaningful (US)	↑ 5.1%*	↑ 8.9%**	↑ 3.9%	↓ 7.7%
Meaningful (India)	↓ 2.3%	↑ 7.0%*	↓ 3.1%	↑ 0.5%
Shredded	↓ 4.0%	↓ 2.8%	↓ 7.2%***	↑ 5.6%
Shredded (US)	↓ 2.3%	↓ 5.0%	↓ 6.1%*	↑ 9.5%
Shredded (India)	↓ 6.8%	↓ 1.6%	↓ 8.7%**	↓ 1.4%

* $p < .05$, ** $p < .01$, *** $p < .001$, black text indicates $p < .10$ and grey text indicates $p > 0.10$

Notes: Rows 1 and 4 consider data from both America and India combined. Columns 1, 2, 3 show the results of regressions and column 4 shows the result of two-sample t-tests. Results reported are from regressions without demographic controls.

Table 2-3: Main treatment effects on quantity of images

	Induced	Induced	Did ≥ 2	Did ≥ 2	Did ≥ 5	Did ≥ 5
Meaningful	0.046*	0.046*	0.047*	0.050*	0.085***	0.088***
	(0.020)	(0.020)	(0.022)	(0.022)	(0.024)	(0.024)
Shredded	-0.040	-0.037	-0.012	-0.005	-0.028	-0.023
	(0.021)	(0.021)	(0.022)	(0.022)	(0.024)	(0.024)
India	-0.185***	-0.183***	-0.170***	-0.156***	-0.035	-0.003
	(0.017)	(0.018)	(0.018)	(0.019)	(0.019)	(0.021)
Male		0.006		-0.029		-0.081***
		(0.018)		(0.019)		(0.021)
Constant	0.848***	0.907***	0.785***	0.873***	0.387***	0.460***
Controls						
Age		0.23		0.29		0.92
Time of Day		0.16		0.06		0.46
Day of Week		0.08		0.00**		0.55
R^2	0.05	0.06	0.04	0.05	0.01	0.02
N	2471	2471	2471	2471	2471	2471

* $p < .05$, ** $p < .01$, *** $p < .001$

Notes: Columns 1, 3 and 5 only include treatments and country. Columns 2, 4, and 6 control for gender, age categories, time of day, and day of week. Rows 6-8 show p -values for the partial F -test for sets of different types of control variables.

Table 2-4: Main treatment effects on quality of images

	Fine Quality					
	Both Countries		United States		India	
Meaningful	0.007 (0.017)	0.014 (0.014)	0.039 (0.023)	0.039* (0.019)	-0.031 (0.025)	-0.013 (0.021)
Shredded	-0.072*** (0.021)	-0.074*** (0.017)	-0.061* (0.027)	-0.066** (0.023)	-0.087** (0.031)	-0.073** (0.023)
India	-0.053*** (0.015)	-0.057*** (0.013)				
Male		0.053*** (0.013)		0.014 (0.017)		0.100*** (0.021)
Labelings 6—10		-0.018** (0.006)		-0.024** (0.008)		-0.016* (0.008)
Labelings ≥ 11		-0.140*** (0.017)		-0.116*** (0.029)		-0.148*** (0.020)
Constant	0.666***	0.645***	0.651***	0.625***	0.634***	0.588***
Controls						
Image		0.00***		0.00***		0.00***
Age		0.10		0.01**		0.25
Time of Day		0.33		0.29		0.78
Day of Week		0.12		0.46		0.26
R^2	0.04	0.15	0.04	0.12	0.02	0.20
N	12724	12724	6777	6777	5947	5947

* $p < .05$, ** $p < .01$, *** $p < .001$

Notes: Robust linear regression clustered by subject for country and treatment on fine quality as measured by the number of cells found two pixels from their exact centers. Columns 1, 3 and 5 include only treatments and country. Columns 2, 4, and 6 control for number of images, the particular image (of the ten images), gender, age categories, time of day, and day of week.

2.6 Appendix

A Detailed Experimental Design

This section details exact screens shown to users in the experimental groups. The worker begins by encountering the HIT on the MTurk platform.

Figure 2-2: The HIT as initially encountered on MTurk



The worker can then click on the HIT and they see the “preview screen” which describes the HIT (not shown) with text. In retrospect, a flashy image enticing the worker into the HIT would most likely have increased throughput. If the worker chooses to accept, they are immediately directed to a multi-purpose page which hosts a colorblindness test, demographic survey, and an audio test for functioning speakers (see Figure 2-3). Although many tasks require workers to answer questions before working, we avoided asking too many survey-like questions to avoid appearing as an experiment.

At this point, the worker is randomized into one of the three treatments and transitioned to the “qualification test.” The page displays an instructional video varying by treatment which they cannot fast-forward. Screenshots of the video are shown in Figures 2-4, 2-5, and 2-6.¹¹

We include the verbatim script for the videos below. Text that differs between treatments is typeset in square brackets separated by a slash. The text before the slash in red belongs to the meaningful treatment and the text following the slash in blue belongs to both the zero-context and shredded treatments.

Thanks for participating in this task. [Your job will be to help identify tumor cells in images and we appreciate your help. / In this task, you'll look at images and find objects of interest.]

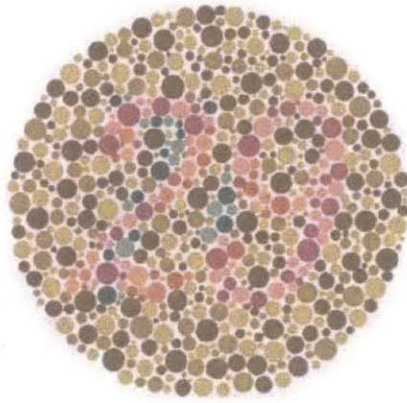
In this video tutorial, we'll explain [three / two] things:

¹¹We thank Rob Cohen who did an excellent job narrating both scripts.


Figure 2-3: The colorblindness test

Since the tasks you will perform require you to be able to differentiate color, we have to ask you a few questions that will determine if you may be colorblind.

1. Look at the below image.



Do you see a number? If so, enter it into this box:

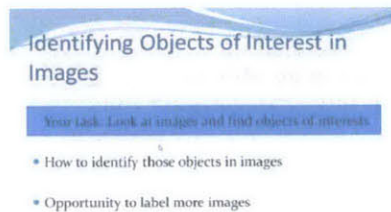
2. Are you male or female?
 Male Female
3. Have you ever had trouble differentiating between reds and greens?
 Yes No
4. Have you ever had trouble differentiating between blues and yellows?
 Yes No
5. How old are you?
6. Listen to the following sound clip () and enter the word below:

Submit

[First, why you're labeling the images, which is to help researchers identify tumorous cancer cells. Next, we'll show you how to identify those tumor cells. / First, we'll show you how to identify objects of interest in images.] [Finally, / Then,] we'll explain how after labeling your first image you'll have a chance to label some more.

Figure 2-4: Opening screen of training video

(a) Zero-context / Shredded treatments



(b) Meaningful treatment

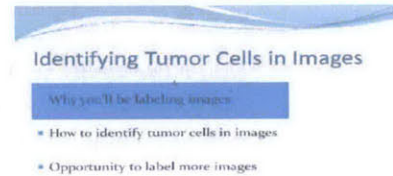
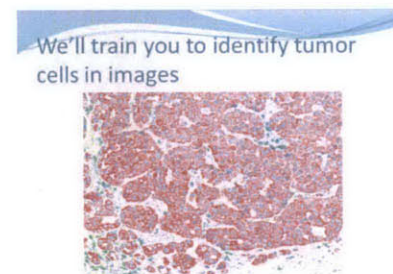
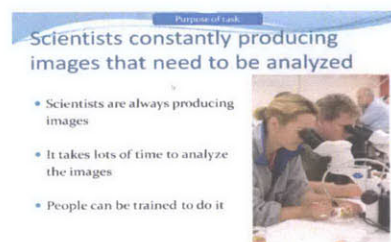


Figure 2-5: Examples of “cues of meaning”



Notes: These cues, which are present in the meaningful treatment, are not shown in the Zero-context or Shredded treatments.

Now we're ready to learn how to identify [tumor cells / objects of interest] in images. Some example pictures of the [tumor cells / objects of interest] you'll be identifying can be found at the bottom left. Each [tumor cell / object of interest] is blue and circular and surrounded by a red border.

When you begin each image, the magnification will be set to the lowest resolution. This gives you an overview of all points on the image, but you'll need to zoom in and out in order to make the most precise clicks in the center of the [tumor cells / objects of interest].

Let's scroll through the image and find some [tumor cells / objects of interest] to identify.

Here's a large cluster of [tumor cells / objects of interest]. To identify them, it is very important to click as closely to the center as possible on each [cell / object]. If I make a mistake and don't click in the center, I can undo the point by right-clicking.

Notice that this [cell / point] isn't entirely surrounded by red, [probably because the cell broke off]. Even though it's not entirely surrounded by red, we still want to identify it as a [tumor cell / object of interest].

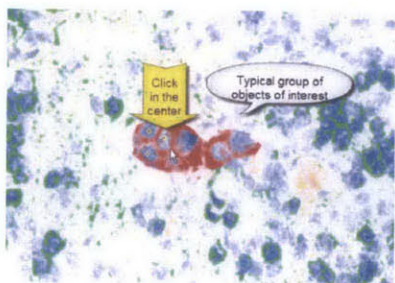
In order to ensure that you've located all [tumor cells / objects of interest], you should use the thumbnail view in the top right. You can also use the magnification buttons to zoom out.

It looks like we missed a cluster of [tumor cells / objects of interest] at the bottom. Let's go identify those points.

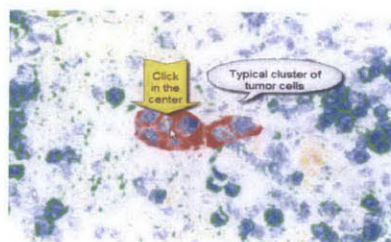
Remember once again, that if you click on something that is not a [tumor cell / object of interest], you can unclick by right-clicking.

Figure 2-6: Illustration of how to perform the task

(a) Zero-context / Shredded treatments



(b) Meaningful treatment



Using the scroll bars, we'll navigate to the other points ... and here's some more to the left ... Now that we think we've identified all points, let's zoom out to be sure and scroll around.

Before submitting, we should be sure of three things: (1) That we've identified all [tumor cells / objects of interest] (2) That we've clicked in the center of each one (3) That we haven't clicked on anything that's not a [tumor cell / object of interest].

Once we've done that, we're ready to submit.

Finally, after you complete your first image, you'll have an opportunity to label additional images as part of this HIT.

The first images you label will pay more to compensate for training.

After that, as part of this HIT you'll have the chance to identify as many additional images as you like as long as you aren't taking more than 15 minutes per image.

Although you can label unlimited images in this HIT, you won't be able to accept more HITs. This is to give a variety of turkers an opportunity to identify the images.

[Thank you for your time and effort. Advances in the field of cancer and treatment prevention rely on the selfless contributions of countless individuals such as yourself.]

Then, workers must take a quiz (see Figure 2-7). During the quiz, they can watch the video freely (which was rarely done).

Upon passing, they began labeling their first image (see Figure 2-8). The training interface includes the master training window where workers can create and delete points and scroll across the entire image. To the left, there is a small image displaying example tumor cells. Above the master window, they have zoom in / out buttons. And on the top right there is a thumbnail view of the overall image.

Participants were given 15 minutes to mark an image. Above the training window, we displayed a countdown timer that indicated the amount of time left. The participant's total earnings was also prominently displayed atop. On the very top, we provided a submit button that allowed the worker to submit results at any time.

Each image had the same 90 cells from various-sized clusters. The cell clusters were selected for their unambiguous examples of cells, thereby eliminating the difficulty of training

Figure 2-7: Quiz after watching the training video (in the meaningful treatment)

Please answer the below questions. Once you answer these questions, you will be qualified to help identify tumor cells.

1 - You should adjust the magnification in order to...

- Make a prettier picture
- Make the tumors exactly 10 pixels across
- Find tumors and make clicks as close to the center as possible

2 - When you are clicking on tumor cells, how many times do you click on the tumor?

- Once
- Twice
- As many dots as you can fit inside the tumor

3 - When you incorrectly click on an area, you should...

- Give up the HIT
- Reload the page
- Use the right mouse button

4 - What will happen if you don't accurately click on the tumor?

- Scientists who are depending on you to identify tumors will not have accurate results
- Your HIT may be rejected
- You will not be allowed to do additional HITs with us
- All of the above.

5 - Your HIT will be rejected if...

- You do not click on all the tumors
- You don't click in the center of the tumor
- You click multiple times on the same tumor or on things that are NOT tumors
- All of the above.

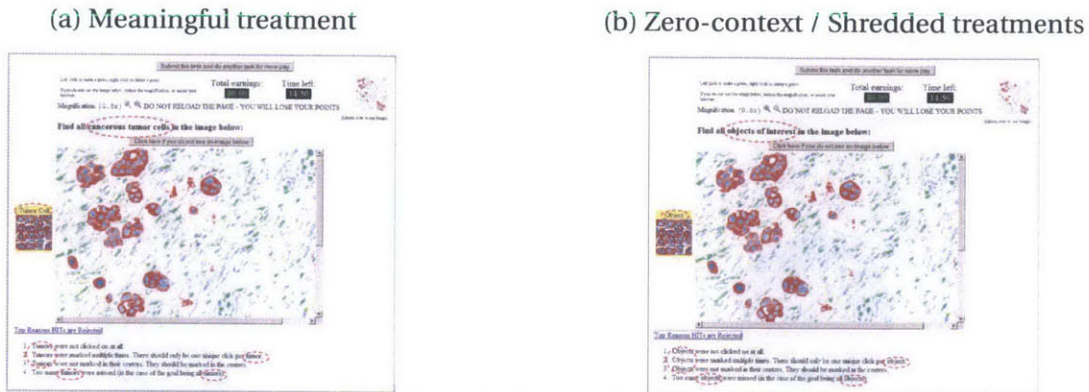
6 - As part of this HIT, how many images will you have the chance to identify assuming you correctly label the tumors?

- As many as you want within a 4hr period (as long as you complete each image within 15 minutes)
- You can label up to 4 more images
- You must submit since you can only label one image

Begin Task

Notes: In the zero-context and shredded treatments, all mentions of "tumor cells" are replaced by "objects of interest." The shredded treatment has an additional question asking them to acknowledge that they are working on a test system and their work will be discarded. Green indicates a correct response; red indicates an incorrect response.

Figure 2-8: Interface for labeling images



Notes: The meaningful interface reminds the subjects in 8 places that they are identifying tumor cells. The zero-context interface only says “objects of interest” and the shredded condition in addition has a message in red indicating that their points will not be saved (unshown). The circles around each point were *not* visible to participants. We display them to illustrate the size of a 10-pixel radius.

the difficult-to-identify tumor cells. In each image, the same clusters were arranged and rotated haphazardly, then pasted on one of five different believable backgrounds using Adobe Photoshop. Those clusters were then further rotated to create a set of ten images. This setup guarantees that the difficulty was relatively the same image-image. Images were displayed in random order for each worker, repeating after each set of ten (repetition was not an issue since it was rare for a participant to label more than ten).

After the worker is finished labeling, the worker presses submit and they are led to an intermediate page which asks if they would like to label another image and the new wage is prominently displayed (see Figure 2-9). In the meaningful treatment, we add one last cue of meaning — a stock photo of a researcher to emphasize the purpose of the task. In the shredded treatment, we append the text “NONE of your points will be saved because we are testing our system, but you will still be paid.” If the worker wishes to continue, they are led to another labeling task; otherwise, they are directed to the post manipulation check survey shown in figure 2-10.

The program ensures that the worker is being honest. We require them to find more than 20% of the cells (the workers were unaware that we were able to monitor their accuracy). If they are found cheating on three images, they are deemed *fraudulent* and not allowed to train more images. Since payment is automatic, this is to protect us from a worker depleting our research account. In practice, this happened rarely and was not correlated with treatment.

Figure 2-9: Landing page after completing a task

(a) Zero-context / Shredded treatments

Thanks for helping to locate the objects of interest in this image! We really appreciate your help.

Your work will be reviewed in the next 15-30 minutes and you will be paid.

You may do another image of similar difficulty especially for you. **Click on the button below:**

Click to do a HIT for \$0.09

Note: You do not have to go through training again

If you are finished, you can click below:

(b) Meaningful treatment

Thanks for helping to locate the tumor cells in this image! We really appreciate your help.

Your work will be reviewed in the next 15-30 minutes and you will be paid.

You may do another image of similar difficulty especially for you. **Click on the button below:**

Click to do a HIT for \$0.09

Note: You do not have to go through training again

If you are finished, you can click below:

Notes: At this point, workers are asked if they'd like to label another image or quit.

Figure 2-10: Post-task survey

Thanks for all of your work. In order to improve our HIT, please complete the following OPTIONAL feedback form.

If you do not want to fill in the survey, please click here:

Please rate how much you agree/disagree with the following statements (where 1-strongly disagree, 5-strongly agree):

The task was fun/enjoyable
 1 2 3 4 5

I liked that the task seemed to be useful and had a good purpose
 1 2 3 4 5

I felt good completing the task
 1 2 3 4 5

The task seemed a lot more meaningful than the average MTurk HITs
 1 2 3 4 5

The task was well-designed and respected my efforts and work more than the average MTurk HITs
 1 2 3 4 5

Any other comments:

Chapter 3

Management and Measurement: Do More Attentive Supervisors Get Better Results?

Abstract

This paper investigates whether greater supervision translates into higher quality work. We analyze data from a firm that supplies answers for one of the most popular question-and-answer (“Q&A”) websites in the world. As a result of the firm’s staffing process, the assignment of supervisors to workers is as good as random, and workers are exposed to supervisors who put forth varying degrees of “effort” (a measure based on a supervisor’s propensity to correct work). Using this exogenous variation, we estimate the net effect of greater supervision and find that a one-standard-deviation increase in supervisor effort reduces the number of bad answers by between four and six percent. By decomposing the total effect into the separate effects on corrected and uncorrected answers, we conclude that supervisor effort tends to *lower* the number of good answers among uncorrected answers. Interestingly, observable worker behaviors (i.e., answer length and time to answer a question) seemed unaffected by supervision. None of the results vary with worker experience.

3.1 Introduction

Most production depends upon the combined efforts of workers and supervisors. Despite the large body of theoretical literature on team production, economists generally lack firm-level data on worker-supervisor interactions. Without such data, we only observe the production output, but we are unable to disentangle the relative contribution of the workers’ “work” and the supervisors’ “supervision.”

Increasingly computer-mediated (Varian, 2010) production environments have made it eas-

ier for economists to observe these interactions, as has the widespread adoption of IT systems that collect detailed, time-stamped metrics for individual workers throughout the production process. Lazear et al. (2011) provides common examples of such jobs — call center workers, technical repair workers, and cashiers. The availability of such data has enabled researchers to answer a variety of questions regarding behavior within the firm.¹

In this paper, we examine how increased supervision affects employee behavior and whether it improves the overall quality of output. More specifically, we measure the effect of being assigned to supervisors who exert greater supervisory “effort”, a measure that is related to each supervisor’s average propensity to correct answers.² Due to the firm’s staffing process, the assignment of supervisors to workers is as good as random and provides exogenous variation in supervisor effort, which allows us to estimate the causal impact of exposure to more proactive supervisors (i.e., ones who exert more effort).

Our study uses production data from an outsourcing company that supplies answers for one of the most popular question-and-answer (“Q&A”) websites in the world. Although these Q&A sites depend on users to answer one another’s questions, the demand for answers routinely exceeds the supply that volunteers provide. To fill this gap, Q&A sites often hire web researchers through outsourcing firms such as the one our study examined.

We observe the “answer production” for over 293,000 questions from July 28th, 2011, until January 23rd, 2012. This entire process takes place on a web-based portal that keeps track of completion times, supervisor’s ratings, and the full text of answers before and after supervisor corrections. Additionally, we observe a quality measure for answers — the number of “thumbs-up” each answer receives from the website’s users. From this we define a “good” answer as one that receives one or more thumbs-up and a “bad” answer as one that receives no thumbs-up while a competing answer submitted by the website’s users does.

We find that primary effect of greater supervision was to reduce low-quality answers. A one-standard-deviation increase in supervisor effort reduced the number of bad answers by between four and six percent. We also find weak evidence that more proactive supervisors reduce the number of good answers.³ Observable worker behaviors, however, did not substantially change; workers did not spend more time answering each question and the length of their answers declined by only one to two percent, an economically insignificant amount. Finally,

¹Lazear (2000)’s analysis of piece-rate and hourly contracts at the SafeLite glass company is a canonical example. More recently, Mas and Moretti (2009) use data from grocery store checkout lines to measure peer effects, Maggio and Alstyne (2011) examine whether access to an information sharing system increased productivity, and Nagin et al. (2002) use a natural field experiment to study opportunistic cheating behavior in call centers.

²See section 3.2.3 for a full definition

³This finding is not robust to the inclusion of worker fixed effects.

these results did not differ across new and more experienced workers.

We also attempt to disentangle the mechanisms by decomposing the net effect of supervisor effort into the separate effects for corrected and uncorrected answers. The main conclusion we draw from this analysis is that supervisor effort tends to *lower* the number of good answers among uncorrected answers. We speculate that this may be a result of workers: 1) “cracking” under pressure, 2) becoming demotivated by having their answers corrected more frequently, or 3) reducing their effort if they expect supervisors will make corrections anyway. No other clear-cut conclusions came out of the analysis and further assumptions would need to be made to place bounds on the magnitude of other effects.

This paper proceeds as follows. Section 3.2 describes the setting of our project and the production process. It also provides a formal definition of supervisor effort and describes it in greater detail. Section 3.3 provides an overview and summary statistics of our data. Section 3.4 describes our empirical strategy and presents results, and Section 3.5 concludes.

3.2 Setting: The firm and production environment

3.2.1 Company background and project description

Our data come from the personnel and productivity records of an outsourcing company that provides services such as transcription, data entry, and web research. We refer to the question-and-answer website as the “client” or “Q&A site” and refer to the outsourcing company as the “firm.”⁴

On the Q&A site, users can ask questions about a wide range of topics. Of more than 20 categories, the most popular are: Science (10%); Computers (8%); Entertainment (7%); and Business, finance, and law (7%).

Figure 3-1 shows an example question from a Q&A site about employment law and whether employers can deduct job expenses from a worker’s paycheck. Answers to the question appear below and, for each answer, website users can write comments and give thumbs-up to helpful answers. We use these thumbs-up as the primary measure of answer quality.

Finally, to the right of each question are sponsored links. Q&A sites receive revenue based on the amount of traffic they generate, not necessarily on the quality of answers. Hopefully, in the case of the legal question, the lawyers who sponsored the links can provide better advice than the users.

⁴At the request of the firm, we do not disclose the name of the Q&A site or firm.

3.2.2 The production process and staffing

The outsourcing firm has several hundred full-time employees who are assigned to one or more projects at a time based on need. The employees who review questions are referred to as “supervisors” or “reviewers” and the employees who answer — but do not review — questions, are referred to as “workers”. Note that supervisors can also answer questions, but their answers still need to be reviewed by different supervisor.

Figure 3-2 shows the production process from when a question is asked until the question is “Removed,” “Expired,” or “Answered/Delivered.”⁵

After a user posts a question, it is sent to the firm. Once it arrives, the question is moderated by a worker who marks it as “valid” or “invalid.” Examples of invalid questions include: advertisements, adult or other inappropriate content, or questions that require extensive research (e.g., “List *all* colleges on the east coast that offer a music major.”)

Next, valid questions are sent to an answering queue where they are researched by the next available worker. Workers also provide a second pass at moderation and can mark questions as invalid.

Once the worker answers a question, it is reviewed by a supervisor who reviews and rates it. The supervisor may then make a final correction to the answer before sending it to the website.

Questions can also “expire” if they are not answered within a specific time frame.⁶ The Q&A website values timely answers and incentivizes the firm to answer them quickly by only paying for on-time answers.

Employees are often assigned to a particular project for the entire shift they work (there is a day shift between 8am and 8pm and a night shift between 8pm and 8am). However, employees may move between the Q&A project and other projects and, in that case, are staffed in half-hour intervals (1:00pm to 1:30pm, etc). Although we do not have precise data on staffing, we impute this and consider an employee to be on duty whenever she answers or reviews a question that was created during that period.⁷

The arrival rate of questions fluctuates unpredictably throughout the day and, since the firm has high-powered incentives to answer questions on time, the firm continually rotates supervisors and workers onto and off of the project in a way that is as good as random. As a result, this provides us with a large amount of exogenous variation in workers’ exposure to supervisors and

⁵Before a question is answered, the website can also “cancel” a question if it’s flagged by users or moderated by the website’s staff; we exclude cancelled questions from our analysis.

⁶From July 28th, 2011 questions expired after eight hours and from November 3rd, 2011, until January 23rd, 2012 (the end of the sample), they expired after one hour.

⁷Given that the intra-day staffing was ad-hoc, this information may never have been recorded.

levels of supervisor effort.

3.2.3 Supervisor effort and its effect on correction rates

This section provides an explicit definition of supervisor effort and describes the assumptions needed for our empirical strategy. Namely, the assignment of supervisors to workers needs to be as good as random and lead to exogenous variation in supervisor effort. Additionally, for supervisor effort to have an effect, workers need to know which supervisors are likely to review their work, and finally, the predicted amount of supervisor effort should correlate with whether a particular worker's answer is corrected.

Definition of supervisor effort

Our measure of supervisor effort, \bar{E}_{wt} , is defined at the worker-shift level based on the aggregate effort put forth by supervisors who are on duty during a shift.

We calculate \bar{E}_{wt} in two steps. First, we define the supervisor effort E_{wst} for each supervisor s relative to a worker w during a shift t . This is done by using each supervisor s 's correction rate for answers submitted by *all other workers* during all shifts *except for* shift t . Second, we obtain \bar{E}_{wt} by averaging the effort of each supervisor on-duty.

Assignment of supervisors to workers

The rotational assignment staffing process leads us to believe that the assignment of supervisors to workers and is essentially random. However, our empirical strategy would be compromised if there were a correlation between supervisor effort and other factors that affect the quality of answers. For example, if supervisors exerted more effort during night shifts and the difficulty of questions varied by shift, our estimates would be biased.

In order to account for this correlation, we add fixed effects for dates and night shift, which weakens the requirement that supervisor effort be exogenous across all periods and only requires that supervisor effort be exogenous within specific periods of time. Given that our sample of supervisors and workers is small, these controls are likely to be important.⁸

Since the firm encourages communication between workers and supervisors, workers generally have an idea of which supervisors are on duty. Supervisors are supposed to answer workers' questions and provide guidance and feedback on their answers. To facilitate this, the firm's

⁸Although the differences in our point estimates with and without controls are not statistically significant in any models, the coefficients sometimes change substantially (see columns 1 and 2 of tables 3-5a, 3-5b, 3-6a, and 3-6b).

IT system provides workers with a screen that shows supervisor ratings⁹ and corrections for each answer along with comments from the particular supervisor who reviewed the answer. Additionally, supervisors and workers are often co-located during a shift.¹⁰

Additionally, the production process ensures that supervisors do not choose which questions they review. After workers answer questions, they go into a queue where they are answered in a first-in, first-out (FIFO) manner by whichever supervisors are on-duty.

3.3 Data and summary statistics

3.3.1 The sample

In our study, we analyze 293,147 questions asked during the period spanning July 28th, 2011, and January 23rd, 2012. Of all questions asked, 56.7% were answered, 32.7% were removed for being invalid,¹¹ and 10.6% expired because they were not answered on-time. In total, we observe answers for 166,246 questions. Although for most analyses, we exclude answers given by supervisors and are left with a sample of 129,874.

The project had two phases: from July 28th, 2011, until November 3rd, 2011, questions expired after eight hours and from November 3rd, 2011, until January 23rd, 2012, questions expired after one hour. During the second period, the Q&A site decided to send a greater volume of questions to the outsourcing firm (2,221 vs. 1,144 per day). We combine these samples since the questions from the site were for all intents and purposes the same.¹²

There were a total of 46 workers who answered 2,811 questions on average and 16 supervisors who reviewed 10,349 answers on average. In table 3-2 and 3-3, we report statistics for workers and supervisors, respectively.

In the average shift, there were 45.5 questions asked, which were answered and reviewed

⁹ These ratings were highly compressed and the firm did not consider them to be reliable indicators of quality; instead of using these to measure quality, we use our external quality metric (i.e., thumbs-up from website users). On a five-point rating scale, 80.7% of answers received a three and 13.8% of answers received a four. Ratings of one, two, and five ratings were only used for 0.7% of answers.

¹⁰ Our measure of supervisor effort would be improved if it incorporated more granular data on the frequency of communication between workers and particular supervisors (i.e., direct conversations, seating arrangements). With this data, we could determine the true amount of supervisor effort each worker faced (by identifying the supervisors to whom they were particular exposed). In addition to increasing precision, this may help our instrumental variables estimation strategy, which we discuss in section 3.2.3.

¹¹ Of these, 69.9% were removed because they required extensive research and 9.8% were removed because they asked for “sensitive” information (e.g., religious beliefs, medical advice, questions of an adult nature).

¹² Additionally, we include time fixed effects for each date to control for variation between the periods.

by an average of 8.9 workers and 2.8 supervisors who were on-duty.¹³ However, the number of questions in a shift was highly variable since the volume depended on the number of website users who asked questions. For example, during the second period, the 25th and 75th percentile shifts had 31 and 82 questions, respectively. The 25th and 75th percentile ranges were 7 and 17 for workers and 2 and 4 for supervisors. Table 3-11 shows further detail on the variability of questions, answers, workers, and supervisors during shifts.

3.3.2 Summary statistics

This section reports summary statistics for our sample. We use separate tables to present summary statistics depending on the levels at which they are defined:

- **Question and answer level** (table 3-1): Fraction of all answers that are corrected
- **Worker level** (table 3-2): Fraction of Worker X's answers that are corrected
- **Supervisor level** (table 3-3): Fraction of answers that Supervisor X corrects
- **Shift level** (table 3-11): Average number of questions asked during a particular shift

We organize our discussion of summary statistics into the following topics: 1) characteristics of answers, reviews, and corrections; 2) supervisor effort and supervisor behavior; and 3) external quality measures.

Answers, reviews, and corrections

Each answer took an average of 6.5 minutes to answer and had a length of 258 characters, while the median answer took 5.9 minutes and had a length of 245 characters (see table 3-1).¹⁴ Although these answers seem short (as a reference, this paragraph is only 475 characters), many answers only require short, factual answers or links to other websites and, therefore the bulk of time is spent researching, not writing. Therefore, an answer may be high quality even if it is short.

There was substantial variability in the total number of answers produced by each worker during the entire project. The average worker submitted 2,811 answers with a large standard deviation of 1,928 (the interquartile range was 1,359 and 4,766). However, this statistic is not

¹³These totals are for both periods. In period one, there was an average of 31.9 questions, 6.2 workers and 2.6 supervisors. In period two, there was an average of 62.3 questions, 12.3 workers and 3.1 supervisors.

¹⁴There are not be practically significant differences in answer lengths across categories. Although there are statistically significant differences in answer length by category ($p < .001$ for an F-test that all differences are zero), the average difference between question categories was less than 25 characters for 26 of the 28 categories (when omitting the category whose average length was closest to the mean).

indicative of productivity per unit of time since employees worker on the project for different lengths of time. Productivity metrics are better analyzed using the amount of time workers took to answer a question. The average worker took 6.8 minutes (sd = 1.3) and the interquartile range was 6.1 minutes and 7.7 minutes. If we translate this into answers per 40-hour work week, the 75th percentile most productive worker produced 393 answers vs. the 25th percentile most productive worker who produced 311 answers.¹⁵ Finally, the average length of an answer for workers did not vary substantially — the average worker’s answer length was 257 characters (sd = 27) and the median was 251. See table 3-2 for more details.

The average question was reviewed by supervisors for 56 seconds and the median review took 30 seconds (table 3-3). Answers that were corrected took more than twice as much time to review as uncorrected ones (88 vs. 42 seconds).

Workers also had their answers corrected at very different rates. The average fraction of a worker’s answers that were corrected was 32.2% with a standard deviation of 8.9% (the interquartile range was 27.3% and 37.0%. The 90th percentile worker had 45.0% of her answers corrected as compared with the median worker whose answers were corrected 31.5% of the time. These and other worker statistics are in table 3-2.

Supervisors corrected 30.9% of answers and, when they corrected an answer, supervisors increased the answer’s length 59.6% of the time (see table 3-1). We also measure how much an answer was changed by its “edit distance” (also known as a Levenshtein distance), which is equal to the number of additions, deletions, or substitutions of characters between the original answer and the corrected answer.¹⁶ The edit distance for corrected answers was just five characters, suggesting that most corrections were superficial spelling or grammatical changes. The average edit distance of corrected answers was somewhat higher, 19.9, and the most-corrected 10 percent of answers had an edit distance of 53 or greater (see table 3-1).

Supervisor effort and supervisor behavior

The number of reviews made by each supervisor varied widely. Although the average supervisor reviewed 10,349 questions, there was large variability in the number of reviews they performed. While the top five supervisors reviewed an average of about 24,000 answers, the bottom five reviewed an average of about 1,000; these “light reviewers” were employees who

¹⁵ Of course, this doesn’t take into account that workers who answered questions faster may have sacrificed quality and done less well of a job.

¹⁶ For example, changing “The Spring flowers **will have** five petals” to “The Spring flowers **have** five petals” represents an edit distance of five since “will” and a space is removed.

primarily answered questions, but were sometimes given supervisory duties.¹⁷

At the question level, there was substantial variability in the amount of supervisor effort a worker experienced while answering each question. For the 129,874 questions the average supervisor effort was 0.312 (sd = 0.084) and the interquartile range of supervisor effort was 0.237 and 0.359 (see table 3-1). Additionally, there was substantial heterogeneity in supervisor effort among supervisors.

The most lenient supervisor corrected only 15.5% of answers, while the most proactive supervisor corrected 49.6% of answers. In terms of edit distance (i.e., the extent of changes), the five most proactive reviewers changed answers by a mean edit distance of 40.3, compared to 15.1 for the five most lenient reviewers. Figure 3-3 illustrates this by plotting each supervisor's average edit distance against the fraction of answers she corrected.

External quality measures (thumbs-up)

To measure the quality of an answer, we collect data on the number of thumbs-up received by each answer — for both our workers and website volunteers. We use these data to define two quality measures: “good” answers and “bad” answers, which we also refer to as high- and low-quality answers. We summarize the overall, *question-level* quality measures in table 3-1 and the quality statistics for *individual workers* in table 3-2.

We define a good answer as one that received one or more thumbs-up. Workers provided good answers for 27.1% of questions¹⁸ and volunteers, collectively, provided good answers for 19.9% of all questions that were answered by the firm's workers.

Unlike our workers, volunteers were not required to answer every question that our workers answered and only answered 51.4% of them. When a question was answered by volunteers, at least one volunteer submitted a good answer 38.3% of the time. However, when just one volunteer submitted an answer, the volunteer's answer was less likely to be good relative to our worker's answer (23.5% vs. 27.1%).

Workers provided bad answers for 8.5% of questions. We define a bad answer as one where the worker's answer received zero thumbs-up, but at least one website user received one or more thumbs-up. This definition captures the idea that the worker could not provide a high-quality answer, but that website users could.

¹⁷Table 3-12 provides more detail on each supervisor and the extent to which they reviewed or answered questions.

¹⁸The vast majority of these, 94.7%, received just one or two thumbs-up; 23.3% received one, 3.9% received two, and only 1.5% received three or more.

At the worker level, the average worker received good answers 27.6% of the time (sd = 3.7%) and the interquartile range was 25.3% and 30.4%. The average worker received bad answers 8.3% of the time (sd = 1.9%) and the interquartile range was 6.8% and 9.5%.

We caution that thumbs-up is not the ideal quality measure since the website values answers based on the revenue they generate, not its users subjective quality rating. However, we use thumbs-up as a proxy since the Q&A site has not given us data on ad revenue and it cannot be collected externally.

3.3.3 Effect of supervisor effort on correction rates for individual workers

In order to ensure that our predicted value of supervisor effort (as defined above) actually influences the amount of supervisory effort faced by each individual worker, we regress an indicator for whether a worker's answer is corrected on the amount of supervision that worker faces during a particular shift.¹⁹

Table 3-4 summarizes the results for this regression and shows that it is robust to the inclusion of dummy variables for dates and night shift, the category of the particular question, and each worker. In all specifications, the relationship is highly significant ($p < 0.001$). We find that a one-standard-deviation increase in supervisor effort causes between a seven and nine percentage point increase in the probability that an answer is corrected (on a base of 30.9%). We also run a balance test by regressing worker experience (i.e., days of work on the project), on supervisor effort (clustering on worker) and find no such correlation ($p = 0.704$).

Finally, we illustrate the strength of this relationship by binning our data and graphing a scatter plot of the probability of correction on supervisor effort (see top panel of figure 3-4). More specifically, we bin supervisor effort into percentiles and calculate the correction rate within each percentile and the average supervisor effort level within each percentile. Finally, we graph the scatterplot of the correction rate vs. the average supervisor effort.²⁰

3.4 Results

This section describes how supervisor effort affected worker behavior and the final quality of answers. Additionally, we break our results down by the experience level of workers.

¹⁹We cluster standard errors on each worker.

²⁰Figure 3-4 shows this scatterplot in the context of illustrating the existence of a first-stage for an instrumental variable regression of answer quality on corrections using supervisor effort as an instrument. However, the same graph also illustrates that predicted supervisor effort correlates with correction rates.

Throughout our analysis, we exclude all employees who, at some point, had reviewed another worker's answers.²¹

3.4.1 Results for worker behavior

Although the most important outcome is the final quality, we are also interested in how greater supervision changes the way workers answer questions.

For each answer, we observe the time workers spend answering a question and the length of their answer. We estimate how this behavior changes as follows

$$B_{qwt} = \alpha + \delta \cdot \underbrace{\bar{E}_{wt}}_{\text{effort}} + \beta X_q + \gamma X_t + \mu X_{wt} + \varepsilon_{qwt} \quad (3.1)$$

where B_{qwt} is the behavior of worker w in answering question q during shift t , which depends on the characteristics of the question X_q , the average supervisor effort a worker experiences during a shift \bar{E}_{wt} , and worker characteristics X_{wt} . Finally, we cluster at the worker level.

Of the controls, X_q includes dummies for the category of a question, X_t includes dummies for the date that a shift occurred and whether it was a night shift, and X_{wt} includes time-varying worker covariates such as experience or time-invariant worker fixed effects.

There are two motivations for using controls. First, given our small data set, they may substantially increase precision. Second, although we argue that the production process results in as-good-as-random assignment, this may be true to a greater extent after controlling for covariates. For example, if night shifts have more proactive supervisors on average and if the difficulty of answers also varies, these controls would be necessary.

Table 3-5a reports a regression of the log time to answer a question on the standardized value of supervisor effort²² and table 3-5b reports the same regression for the log length of an answer. Rather than levels, we use a log specification since, as shown in table 3-1, both of these variables are highly skewed.

For tables 3-5a and 3-5b, column 1 estimates a model without any controls, column 2 includes dummies for each individual date and whether a shift takes place in the evening (i.e., a

²¹Although several employees could potentially be classified as a worker or supervisor, any employee who made a review is not counted as a worker. Table 3-12 lists all reviewers and classifies them as either "heavy" or "light" supervisors depending on whether they reviewed more answers than they answered questions. We report our results excluding both types, but including the light reviewers doesn't change results.

²²We standardize supervisor effort using the mean (31.2%) and the standard deviation (9.0%) of supervisor effort defined at the worker-shift level for all workers and shifts.

night shift), column 3 adds controls for the type of question, and column 4 contains all of the aforementioned controls, plus fixed effects for each worker.

Under none of the specifications do we find any relationship between effort and time to answer a question. We do see, however, that including dummies for date and night shift increase the R^2 from practically zero to just above 0.05, suggesting that there are important differences over time and between categories. Not surprisingly, including fixed effects for each worker improves the R^2 to 0.153, which emphasizes the importance of worker heterogeneity. The results for answer length follow the same pattern.

Although the time to answer a question was unchanged, workers produce shorter answers when under greater supervision. A one-standard-deviation increase in supervisor effort leads to an approximately two percent decrease in the length of answers with and without controls for date, night shift, and category. After adding fixed effects, the effect is reduced to about a one percent decrease.

Although these results are significant at the 1 percent level, they are probably too small to be economically meaningful — the reduction is only about five characters.²³

3.4.2 Results for answer quality

Although worker behaviors (i.e., time to answer a question and answer length) did not substantially change, the quality of answers still could. For example, if more proactive supervisors made improvements to work, quality would go up even if workers did not change their behavior. It is also possible that worker effort increased, but that our observables serve as poor proxies.

As before, we use the below equation to estimate the effect of supervision on two measures of quality: “good” answers and “bad” answers.²⁴

$$Y_{qwt} = \alpha + \delta \cdot \underbrace{\bar{E}_{wt}}_{\text{effort}} + \beta X_q + \gamma X_t + \mu X_{wt} + \varepsilon_{qwt} \quad (3.2)$$

where Y_{qwt} is a quality of a question answer by worker w during shift t , which depends on the characteristics of the question X_q , the average supervisor effort a worker experiences during a shift \bar{E}_{wt} , and worker characteristics X_{wt} . Notably, even though answers may be corrected by

²³Based on a two percent reduction on an average answer length of 258 characters.

²⁴As described in section 3.3, a good answer is one where the firm’s worker provided an answer that received one or more thumbs-up from website users. A bad answer is one where the worker’s answer did not receive any thumbs-up, but where one or more volunteers received one or more thumbs-up.

supervisors, we do not index questions by supervisors since we are estimating the net effect of the *average* supervisor effort faced during a shift.²⁵ As before, we cluster at the worker level.

Table 3-6 reports estimates for how supervisor effort affects quality. We find strong evidence that greater supervision reduced the number of bad answers and mixed evidence that it increased the number of good answers.

A one-standard-deviation increase in supervisor effort increased the number of good answers by 0.84 percentage points in the specification without any controls. Adding controls for date, night shift, and question category yielded smaller estimates closer to one-half of a percentage point, which represents a decrease in high-quality answers of about two percent ($p < 0.05$).²⁶ However, after adding fixed effects for each worker, there is no statistically significant effect of supervision on high-quality answers and, with 95% confidence, we can rule out effects as larger than a one percent increase or a two percent decrease.²⁷

Unlike the results for high-quality answers, the reduction in low-quality answers is larger and more robust. Specifications with and without controls yield point estimates around -0.005 , suggesting that one-standard-deviation more supervision reduces the number of bad answers by 6.0%.²⁸ After adding fixed effects for workers, the coefficient declines to -0.0033 , which still represents about a four percent reduction in bad answers. All of these estimates are significant at the 1% level.

These results suggest that supervisor effort weeds out bad answers, but has less of an effect on improving already-acceptable answers.

Additional measures of performance

We consider two additional outcomes that relate to performance, but have more ambiguous interpretations than the quality measures: namely, the probability that no volunteers answer the question and the probability that a question expires. Table 3-7 reports estimates of the effect of supervisor effort on both of these outcomes.

Volunteers' may decide to answer a question for a variety of reasons. What can we infer about the quality of a worker's answer if, after the worker submits a response, no one volunteers to provide another answer?²⁹

On the one hand, it could be a positive signal if the worker's answer was high quality and it

²⁵And, the average supervisor effort depends on the effort of *all* supervisors on a shift.

²⁶A 0.5 percentage point decrease on a base of 27.1% \approx a 1.8 percent decrease in good answers.

²⁷With $\beta = -0.00140$ and a standard error of 0.00224, the confidence interval is $(-0.0214, 0.0110)$, which, on a base of 10.6% corresponds to an increase of 1.1% and a decrease of 2.1%.

²⁸A 0.5 percentage point decrease on a base of 8.4% \approx 6.0 percent decrease in good answers.

²⁹We only observe volunteer answers for questions where our workers submitted answers

fully answered the question or crowded out other answers from volunteers. On the other hand, it would be a neutral signal if the question was not interesting, poorly specified, or otherwise undesirable to answer. If supervisor effort were uncorrelated with question characteristics, the lack of volunteer answers may indicate a good worker response. However, the interpretation would be complicated if supervisor effort also affected which questions were moderated and marked as invalid.

Table 3-7a we present results of a regression of volunteer non-response on supervisor effort. In all specifications, there is a significant effect at the 1% level. For the unadjusted regression and those with dummies for date, night shift, and question category, a one-standard-deviation increase in supervisor effort increased non-response by approximately two percentage points, which represents about a four percent increase in volunteer non-response (on a base of 48.6%). After adding worker fixed effects, the effect drops to about one percentage point (an approximately two percent increase in volunteer non-response). If we choose to interpret this as a positive signal, this provides some evidence that supervisor effort improves answer quality.

Finally, table 3-7b examines whether the probability that an answer expires depends on the overall supervisor effort during a shift. For example, supervisors who exert more effort may cause workers to answer questions more quickly. This would be consistent with the finding that workers answers become shorter while working for more proactive supervisors. Our regression uses all questions that were answered, expired, or marked as invalid and includes dummies for date and night shift. We cannot include worker fixed effects since expired or moderated questions are not always assigned to workers and, because our variation in supervisor effort occurs at the shift level and is not specific to any workers, we cluster on shifts.

In none of these specifications to we find a statistically significant effect of supervisor effort on whether a question expires. However, these estimates are extremely imprecise.

Heterogeneity by experience

Finally, we test for heterogeneous effects on quality depending on a worker's experience. We classify a worker as "new" if she answered fewer than 500 questions at the time she submitted an answer.³⁰ Under this definition, 16.4% of questions were answered by new workers.

It is not clear how supervisor effort should interact with experience. New workers may already be under greater supervision than would be suggested by the supervisor effort variable,

³⁰500 questions corresponds to a little more than 50 hours of work. Other cutoffs such as 250 or 1,000 questions were considered and did not substantially change results. As the experience threshold is lowered, the effect of inexperience may be higher, but becomes difficult to detect due to the small sample size. As the threshold is raised, there is less reason to expect experience effects to be important and there will be fewer workers in our sample.

which would mute any effect. On the other hand, if more proactive supervisors were more likely to give feedback to new workers, the quality of new workers' answers would be very responsive to supervisor effort. Another consideration is that if workers learn how to game the system over time, the quality of their answers would be more responsive to supervision if they shirk more when they aren't being observed.

In table 3-8, we estimate whether worker experience interacts with supervisor effort by adding interaction terms and dummies for whether a worker is new. All specifications include controls for date, night shift, and question category, while some specifications also include worker fixed effects.

Columns 2 and 3 and columns 5 and 6 include a dummy for whether a worker is new and interacts that dummy with supervisor effort. The coefficient on the new worker dummy identifies whether new workers are more likely to provide better answers. The coefficient on *New worker* \times *Supervisor effort (std)* tests whether inexperienced workers respond differently to supervision. In none of these specifications do we find a differential effect of supervisor effort.³¹ However, the estimates are imprecise and we may be underpowered given our sample size. Interestingly, none of the coefficients on *New worker* are significant at the 5 percent level, indicating that experience has no effect on quality.³²

In summary, we find little evidence that supervision has a different effect on more experienced workers.

3.4.3 Mechanisms for how supervision affects quality

The estimates presented thus far show the net impact of supervision. However, they are not informative about the channels through which supervisors might improve output. In particular, we consider the following two channels:

1. **Monitoring effect:** how increased supervision affects the quality of a worker's answer
2. **Correction effect:** how supervisor corrections change the quality of a corrected answer

To try to disentangle these effects, we use two strategies. First, we decompose the total quality effect into separate quality effects for corrected and uncorrected answers. Second, we estimate the causal effect of a supervisor's corrections on quality using supervisor effort as an instrument for whether a correction is made. However, we view these results skeptically since

³¹The lowest p-value is $p = .495$

³²The coefficient in column 6 is the closest one to statistical significance ($p = 0.110$). Again, this may be a result of being underpowered rather than the absence of an effect.

the exclusion restriction is probably not satisfied since it would require that worker effort be unaffected by supervisor effort, which seems unlikely.

Decomposition of quality effects

In order to understand the mechanisms through which supervisor effort affects quality, we decompose the causal effect of supervisor effort on the quality of an answer.³³ More specifically, we estimate the following equations

$$QUALITY_{qwt} = \alpha_1 + \beta_1 EFFORT_{wt} + \varepsilon_{qwt} \quad (3.3a)$$

$$QUALITY_{qwt}^C = \alpha_2 + \beta_2 EFFORT_{wt} + \varepsilon_{qwt} \quad (3.3b)$$

$$QUALITY_{qwt}^{NC} = \alpha_3 + \beta_3 EFFORT_{wt} + \varepsilon_{qwt} \quad (3.3c)$$

where $QUALITY_{qwt}^C$ ³⁴ represents an indicator for whether an answer was good or bad times an indicator for whether an answer was *corrected* and where $QUALITY_{qwt}^{NC}$ represents an indicator for whether an answer was good times an indicator for whether an answer was *not corrected*. Definitionally, it must be true that $\beta_1 = \beta_2 + \beta_3$.

We report the results of these regressions in table 3-9. In the *Outcomes* panel, we list estimates for β_1 , β_2 , and β_3 are shown in columns 1, 2, and 3, respectively. Separate estimates are reported for the outcomes of good answer and bad answer. In the second panel, we show how the magnitude of the quality change compares to the fraction of corrections that are made. More specifically, we divide each coefficient in the *Outcomes* panel by the coefficient in the *Regression of Corrected on Effort* row.

Before moving to interpretation, we highlight the existence of a **mechanical effect** — i.e., more effortful supervisors *mechanically* raise the average quality of corrected and uncorrected answers. This is because higher-effort supervisors begin making corrections by drawing from the lowest-quality answers from the previously uncorrected answers, which will raise the quality of both the corrected and uncorrected answers.³⁵ This fact is important since it helps us sign and interpret the coefficients on our decomposition.

³³One caveat regarding this decomposition is that the supervisor's correction is endogenous.

³⁴Subscripts are defined as follows: q is a question answer, w is a worker, and t is a particular shift. As described in section 3.2.3, the supervisor effort variable lives at the worker-shift level.

³⁵This phenomena is known as the "Will Rogers effect" based on a joke he told, "When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states." In this case, the phenomena occurs because the Okies are below average intelligence within Oklahoma, but above average intelligence within California.

First, we consider the effect on good answers. Column 3 of the “Good answer” row from the *Outcomes panel* in table 3-9 shows that the coefficient on uncorrected answers (β_3) is negative (-0.0245) indicating that uncorrected answers actually got worse. This means that the effect on the quality of uncorrected answers was sufficiently negative that it reversed the mechanical effect, which would have led to a positive coefficient. There are two possible explanations: 1) supervisors tended to improve answers that were already good, or 2) workers provided worse answers under greater supervision. We consider it more likely that supervisors triaged bad answers than tried to further improve already good answers, therefore this suggests that the quality of workers’ answers decreased under greater supervision. Some possible explanations for this are that workers: 1) “crack” under pressure, 2) become demotivated by having their answers corrected more frequently, or 3) reduce effort if they expect their supervisors will make corrections anyway.

From the positive coefficient on the corrected answers, .0231, we are unable to make any firm conclusions. The mechanical effect would also make this coefficient positive; however, we don’t know whether it is more positive than it otherwise would have otherwise been and would need to make further assumptions in order to put bounds on the effects.

For the bad answers, the mechanical effect works in the opposite direction and tends to *lower* the quality of both corrected and uncorrected answers. Column 3 of the “Bad answer” row indicates the coefficient on uncorrected bad answers is -0.0111, which suggests that uncorrected answers became less “bad” (i.e., they improved). However, the mechanical effect would have made the coefficient negative absent any effect. Therefore, as in the case of corrected good answers, we cannot draw any conclusions without further assumptions.

The second panel of table 3-9 shows the size of quality changes relative to the reduced form effect of supervisor effort on the fraction of answers that were corrected — the RF effect is 0.0895 and the “reduction in badness” is -0.0034. Since this reduction in badness is so small relative to the fraction of corrections made ($0.038 = \frac{-0.0034}{0.0895}$), we see that, even in the best case, corrections only serve to make the answers less bad and cannot have that large of an effect. The remaining ratios in panel 2 can be used to perform other bounding exercises.

Instrumental variables estimation

We could potentially estimate the causal effect of a supervisor’s correction on quality if we had an instrument that affected whether a correction is made. One possible instrument is to use the propensity of a supervisor to make corrections to other people’s work (i.e., supervisor

effort) as an instrument for whether a particular worker's answers are corrected.³⁶ However, since workers probably adjust effort depending on supervisor effort, the exclusion restriction is likely to be violated. Moreover, the direction of bias is non-obvious. On the one hand, effort may increase if more proactive supervisors penalize workers. On the other hand, worker effort could decrease if workers expect supervisors will correct their answers.

In table 3-10, we demonstrate the existence of a strong first stage and also present the IV and reduced form estimates. Figure 3-4 presents these results in the form of a visual IV. Although there is a strong first stage ($p < 0.001$), we remain skeptical of the IV results since the exclusion restriction is probably violated.

The instrumental variable estimate for the effect of corrections on *good* answers is not statistically significant after including worker fixed effects (column 4 of table 3-10).³⁷ On the other hand, the IV estimates for the probability of having a *bad answer* are all highly significant ($p < 0.01$) and suggest that a supervisor who corrects all answers vs. one who corrects no answers reduces the probability of a bad answer by approximately four percentage points in the specification with worker fixed effects (column 4), a reduction of 48 percent (on a base of 8.4%).

Finally, we consider how a violation of the exclusion restriction would bias our estimates. The most likely source of bias is that increased supervision raises worker effort, which would reduce the number of bad answers. If this were true, the effect of higher worker effort would load onto the IV estimates and make them larger (in absolute value) than they should be. Additionally, column 3 of table 3-9 (in the first panel) shows that the coefficients on uncorrected answers are strongly significant. This suggests that supervisor effort affected the behavior of workers, which lends further support to the idea that the exclusion restriction has been violated. To summarize, although we report the IV estimates, we hesitate to draw any conclusions from them.

3.5 Conclusion

Our paper analyzes whether increased supervisor effort achieves better results. We use data from an outsourcing firm whose staffing process resulted in as-good-as-random assignment of

³⁶Similar instruments are used by Doyle (2007), who use the propensity of child protection investigators to place children in foster care in order to measure the effect of placement and by Kling (2006), who uses judges' propensity to deliver harsher sentences in order to measure the effect of incarceration length.

³⁷Under other specifications (columns 1 to 3), the coefficient on the IV estimates suggests that a supervisor who corrects all answers vs. one who corrects no answers decreases the probability of having a good by approximately 6 to 10 percentage points ($p < 0.05$), a reduction of between 21 and 28 percent (on a base of 28.7%). However, given the amount of heterogeneity of workers, we do not believe that models without worker fixed effects are credible.

supervisors to workers. As a result, this created exogenous changes in supervisors (and levels of supervision) and allowed us to observe how workers responded to different levels of supervision.

The net effect of greater supervision was to reduce the number of low-quality answers; a one-standard-deviation change in supervisor effort reduced bad answers by between four and six percent. There was also limited evidence that more supervision reduced the number of good answers, but this was not robust to the inclusion of worker fixed effects. Additionally, increased supervision did not have a significant effect on observable worker behavior (i.e., answer length and time to answer a question). Finally, the impact of greater supervision seemed the same irrespective of worker experience.

We attempt to disentangle the mechanisms by decomposing the net effect of supervisor effort into the effect on corrected and uncorrected answers. The main conclusion we draw from this analysis is that supervisor effort tends to *lower* the number of good answers among uncorrected answers.³⁸ We speculate that this may be a result of workers: 1) “cracking” under pressure, 2) becoming demotivated by having their answers corrected more frequently, or 3) reducing their effort if they expect supervisors will make corrections anyway.

The most severe limitation of our study is that cannot precisely identify the channels through which supervision matters. In future research, one could run an experiment that exogenously varies whether supervisors correct work and whether workers believe their work will be reviewed. Such a design would help distinguish among the different theories that could explain our results.

³⁸In general, we are unable to separately identify the correction effect, the monitoring effect, and the mechanical effect.

Figure 3-1: Example question from a Q&A website

The screenshot displays a Q&A website interface. At the top, a navigation bar shows "All > Business, Finance & Law". The main content area features a question posted by user "lorenzofaure:" 10 hours ago via iPhone. The question is: "Can my employer deduct money from my check for gas if he provides transportation to a far away jobsite?". The question text states: "The jobsite is located more than 100 miles from our town. We're required to be at a certain place every morning where a company vehicle pick us up and take us to the workplace. The employer has been taking money out of our checks to account for gas." Below the question is an "Answer This Question" button, a "Report as" dropdown, and social media sharing options for "Like" and "Tweet".

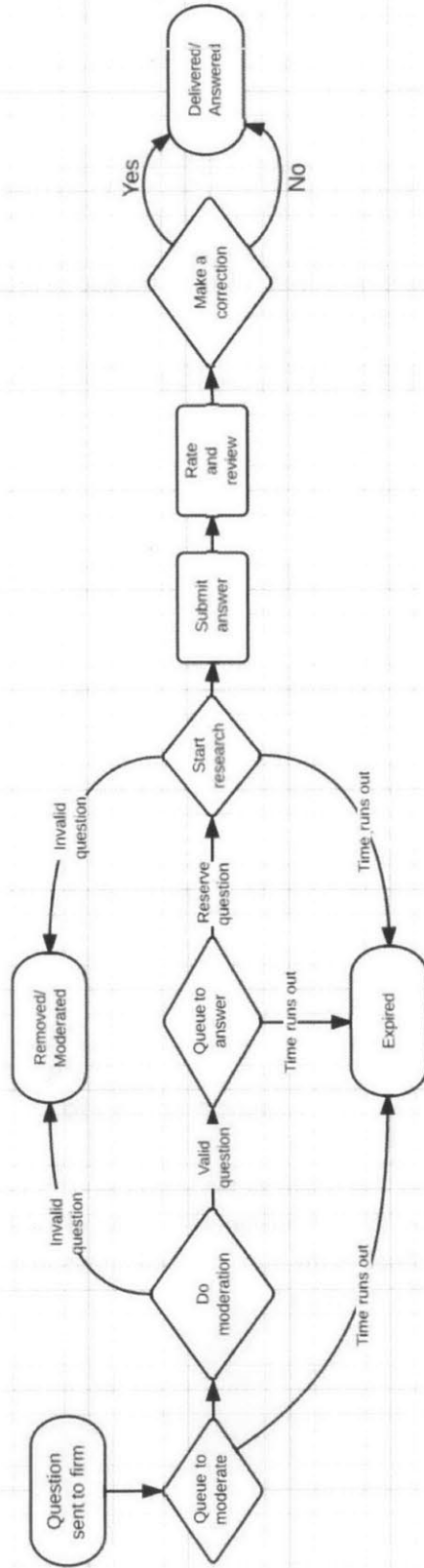
Three answers are provided:

- gatorblu:** 9 hours ago. Answer: "That doesn't sound right. Contact your state's labor board." Interactions: 5 Helpful, Fun, Comments (0), Report as.
- medic614:** 9 hours ago. Answer: "I wouldn't think so; but if it is, you should get a receipt so you can deduct it on your income taxes." Interactions: 4 Helpful, Fun, Comments (0), Report as.
- fossil5400:** 9 hours ago. Answer: "That absolutely sounds wrong, you probably should not ask your employer, or, if you do, go through your Human Relations Dept." Interactions: Helpful, Fun, Comments (0), Report as.

On the right side, a "Popular Searches" section lists various topics such as "Florida Labor Laws", "Salaried Employee Rules", "Unlawful Deduction from Wages", "Can My Employer Withhold Money from My Paycheck?", "Can My Employer Change My Pay?", "Employer Deductions from Pay", "Can an Employer Take Money from Your Check?", "Ohio Labor Laws", "Employee Rights Termination", and "Employee Rights". Below this is a red bracketed area labeled "Links to advertising".

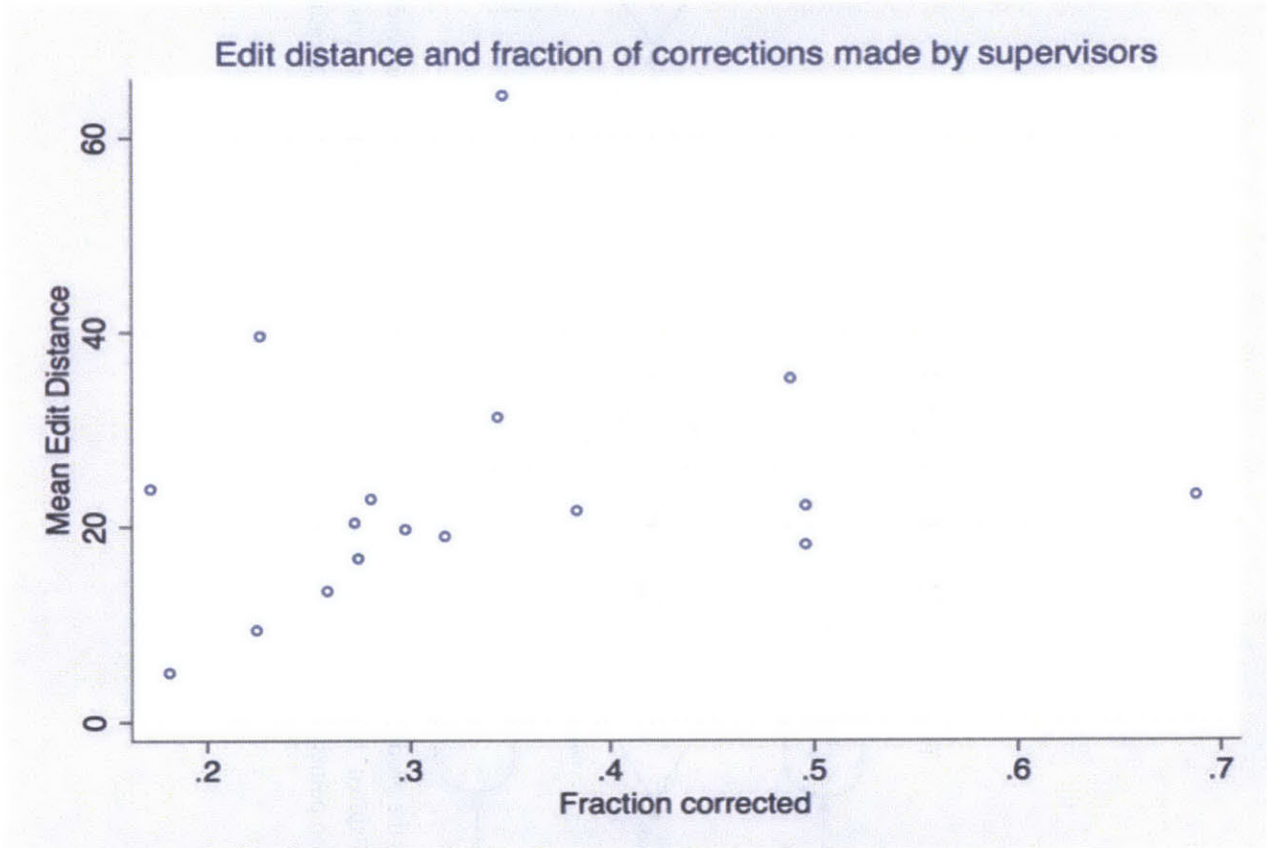
Notes: This figure shows an example question posted to a Q&A site along with several answers.

Figure 3-2: Flowchart of production process



Notes: The figure provides an overview of the production process. After a question is asked, it can either be “Removed” through moderation (e.g., if the question asks about an illegal activity), “Expired” if a question is not answered on-time, or “Delivered” if a question is answered and reviewed within the time limit. Diamonds indicate decision points, rectangles represent activities, and ovals indicate initial and terminal nodes.

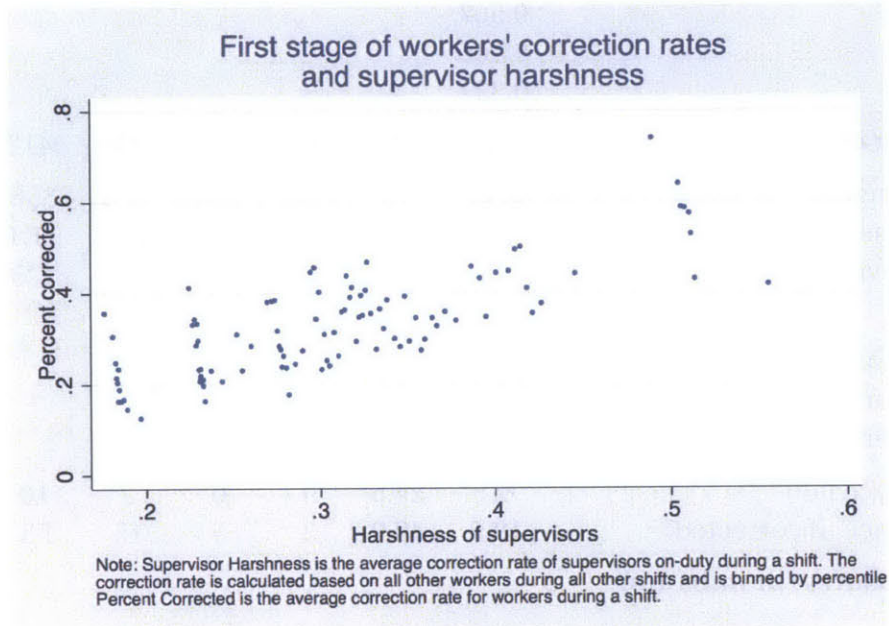
Figure 3-3: Heterogeneity in supervisor effort and the average size their corrections



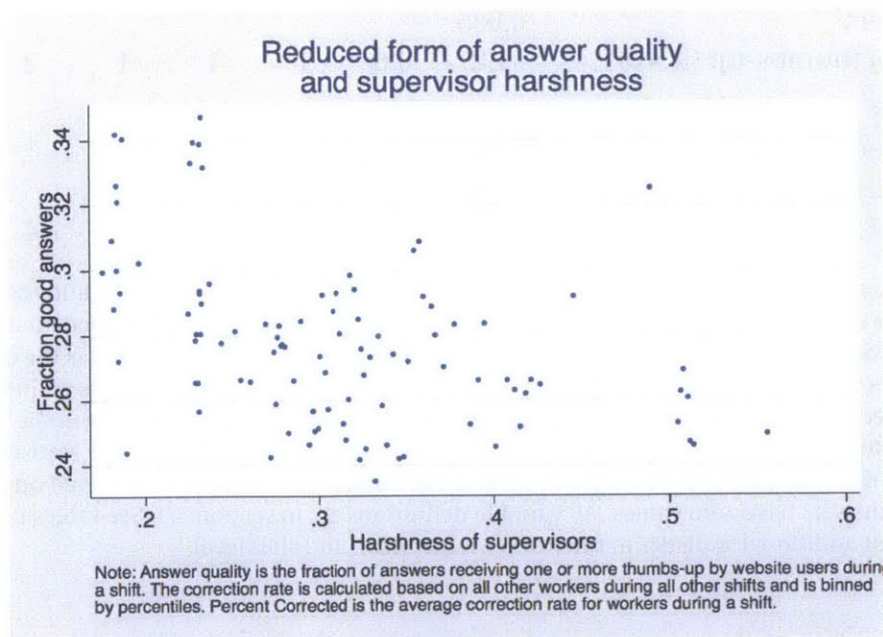
Notes: Each point represents a supervisor. The x-axis shows supervisor effort (i.e., the fraction of answers each supervisor corrected) and the y-axis show the “edit distance” (i.e., size of a change) for when the supervisor corrected an answer. Edit distance is defined as the number of additions, deletions, or substitutions of characters between an initial answer and the corrected answer.

Figure 3-4: Visual IV: Supervisor effort, correction rate, and answer quality

(a) First stage of correction rate and supervisor effort



(b) Reduced form of answer quality and supervisor effort



Notes: The upper figure shows the first-stage relationship for the fraction of a worker's question-answers that are corrected (during a shift) on supervisor effort. The lower figure shows the reduced form relationship for the fraction of high-quality (e.g., "good") answers on supervisor effort.

Table 3-1: Summary statistics: Questions and answers

Variable	Mean	SD	25th	50th	75th	90th	N
Questions & answers							
Answered	0.567						293,147
Removed	0.106						293,147
Expired	0.327						293,147
Supervisor effort	.312	.084	.237	.307	.359	.412	129,874
Answer length	258	75	207	245	296	356	129,874
Time to answer (mins)	6.5	3.3	4.2	5.9	8.2	11.1	129,874
Time to review (secs)	56	78	18	30	62	125	129,874
Corrections							
Fraction corrected	0.309						129,874
Fraction lengthened (all)	0.184						129,874
Fraction lengthened (if corrected)	0.596						40,102
Edit distance (all)	6.1	24.5	0	0	2	10	129,874
Edit distance (if corrected)	19.9	40.9	2	5	15	53	40,102
Quality measures (thumbs-up)							
<u>Workers</u>							
Good answers	0.271						129,327
Bad answers	0.085						129,327
Number of thumbs-up (if > 0)	1.2	0.6	1	1	1	2	35,056
<u>Volunteers</u>							
% Answered by volunteer	0.514						129,327
% Good answer (if answered)	0.383						66,487
Number of volunteers (if > 0)	2.2	2.2	1	2	3	4	66,487

Notes: This table presents summary statistics for the 293,147 questions that were asked and the 129,874 answers provided by 46 workers. Although we observe 166,246 answers, we exclude 36,372 answers that came from 16 supervisors (leaving 129,874 answers). All answers were also reviewed. *Supervisor effort* is the average effort level of all supervisors on duty while a worker answered each question. *Fraction lengthened* is an indicator variable for whether a worker's answer was lengthened by a supervisor's correction. *Fraction lengthened* and *Edit distance* are reported separately both for the full sample and for corrected answers only. When these statistics are calculated only on corrected answers, the un-corrected answers have missing values; when calculated on all answers, the non-corrected answers have zero values. All variable definitions are in section 3.3. See tables table 3-2, table 3-3, and table 3-11 for additional statistics at the worker, supervisor, and shift level.

Table 3-2: Summary statistics: Worker behavior

Variable	Mean	SD	25th	50th	75th	90th	N
Answers							
Avg. number of answers	2,811	1,928	1,359	2,428	4,766	5,776	46
Avg. time to answer (mins)	6.8	1.3	6.1	6.7	7.7	8.5	46
Avg. answer length	257	27	238	251	267	307	46
Corrections received							
Avg. fraction corrected	0.322	0.089	0.273	0.314	0.37	0.45	46
<u>Stats if corrected</u>							
Avg. % of answers lengthened	0.593	0.064	0.549	0.597	0.629	0.662	46
Avg. edit distance	20.3	6.5	15.1	21.1	23.8	26.4	46
<u>Stats if uncorrected</u>							
Avg. % of answers lengthened	0.192	0.058	0.143	0.191	0.225	0.274	46
Avg. edit distance	6.8	3.4	4.3	6.4	9.1	11.3	46
Quality of answers submitted							
Avg. % good answers	0.276	0.037	0.253	0.271	0.304	0.318	46
Avg. % bad answers	0.083	0.019	0.068	0.08	0.095	0.105	46

Notes: This table provides summary statistics for the 46 workers who answered 129,327 questions. Unlike table 3-1, which describes statistics for all questions and answers, this table describes the distribution of *each supervisor's* answering behavior. For example, the row *Avg. number of answers* says that the median worker (of 46 workers) answered 2,428 questions. All variable are defined in section 3.3.

Table 3-3: Summary statistics: Supervisor behavior

Variable	Mean	SD	25th	50th	75th	90th	N
Reviews							
Avg. number of reviews	10,349	9,585	1,093	8,721	16,766	26,999	16
Avg. time to review (secs)	66	21	49	62	86	99	16
Corrections given							
Supervisor effort (avg. fraction of answers corrected)	0.348	0.134	0.267	0.308	0.437	0.497	16
<u>Stats if corrected</u>							
Avg. % of answers lengthened	0.582	0.086	0.503	0.563	0.645	0.715	16
Avg. edit distance	24.9	12.9	18.5	21.8	27.3	39.1	16
<u>Stats if uncorrected</u>							
Avg. % of answers lengthened	0.203	0.094	0.15	0.169	0.235	0.31	16
Avg. edit distance	8.8	5.5	5.1	7.3	10.8	17.2	16
Quality of answers reviewed							
Avg. good answers	0.292	0.019	0.282	0.289	0.302	0.323	16
Avg. bad answers	0.077	0.018	0.065	0.071	0.093	0.104	16

Notes: This table provides summary statistics for the 129,327 answers that were reviewed by all 16 supervisors. Unlike table 3-1, which describes statistics for all questions and answers, this table describes the distribution of *each supervisor's* reviewing behavior. The *Quality of answers reviewed* panel evaluates the average quality of answers. For example, the row *Avg. number of review* says that the average supervisor (of 16 supervisors) reviewed 10,349 questions. Table 3-12 shows a list of all reviewers and their number of reviews and supervisor effort. All variable are defined in section 3.3.

Table 3-4: Effect of predicted supervisor effort on corrections

	Outcome = Answer is corrected			
	(1)	(2)	(3)	(4)
<i>% Answers corrected</i>	0.309			
Supervisor effort (std)	0.0818*** (0.00890)	0.0724*** (0.00895)	0.0727*** (0.00893)	0.0895*** (0.00712)
Fixed effects				
Date and nightshift		x	x	x
Question category			x	x
Workers				x
N	129,327	129,327	129,327	129,327
# Workers	46	46	46	46
# Shifts	6,283	6,283	6,283	6,283
R^2	0.0274	0.0451	0.0461	0.0855

Robust standard errors clustered by worker. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table tests whether the predicted amount of supervisor harshness that a worker faces during a shift is correlated with whether that worker's answer is corrected. The unit of observation is one of 129,327 answers that were provided by 46 workers during 6,283 shifts. The row *Supervisor effort (std)* reports the coefficient from a regression of whether a worker's answer is corrected and the standardized amount of supervisor effort faced by the worker during the shift. Column 1 reports the unadjusted regression and columns 2-4 include various controls for time, type of question, and individual workers. We standardize supervisor effort based on the average supervisor effort faced by all workers during all of their shifts.

Table 3-5: Effect of supervisor effort on worker behavior

(a) Time workers take to answer a question

	Outcome = Log time to answer			
	(1)	(2)	(3)	(4)
Supervisor Effort (std)	0.00468 (0.0161)	-0.0236 (0.0160)	-0.0244 (0.0159)	-0.00128 (0.00542)
Fixed effects				
Date and nightshift		x	x	x
Question category			x	x
Workers				x
N	129,327	129,327	129,327	129,327
R ²	0.0000468	0.0523	0.0546	0.153

Robust standard errors clustered by worker. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(b) Length of a worker answers

	Outcome = Log character length of answer			
	(1)	(2)	(3)	(4)
Supervisor effort (std)	-0.0209*** (0.00529)	-0.0199** (0.00657)	-0.0195** (0.00647)	-0.00835** (0.00303)
Fixed effects				
Date and nightshift		x	x	x
Question category			x	x
Workers				x
N	129,327	129,327	129,327	129,327
R ²	0.00458	0.0345	0.0437	0.151

Robust standard errors clustered by worker. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table reports the effect of supervisor effort on worker behavior. Each observation represents an answer to one of 129,327 questions answered by 46 workers during 6,283 shifts. The upper panel shows the effect of supervisor effort on the amount of time workers spend answering a question and the lower panel shows the effect of supervisor effort on the length of an answer.

Table 3-6: Effect of supervisor effort on answer quality

(a) Probability that answer is high quality

	Outcome = "Good" answer (mean = 0.271)			
	(1)	(2)	(3)	(4)
Supervisor effort (std)	-0.00844** (0.00267)	-0.00591** (0.00219)	-0.00485* (0.00210)	-0.00140 (0.00224)
Fixed effects				
Date and nightshift		x	x	x
Question category			x	x
Workers				x
N	129,327	129,327	129,327	129,327
R ²	0.000316	0.0130	0.0229	0.0254

Robust standard errors clustered by worker. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(b) Probability that answer is low quality

	Outcome = "Bad" answer (mean = 0.085)			
	(1)	(2)	(3)	(4)
Supervisor effort (std)	-0.00556*** (0.00124)	-0.00502*** (0.00137)	-0.00459*** (0.00129)	-0.00338** (0.00103)
Fixed effects				
Date and nightshift		x	x	x
Question category			x	x
Workers				x
N	129,327	129,327	129,327	129,327
R ²	0.000348	0.00630	0.0142	0.0156

Robust standard errors clustered by worker. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table reports the effect of supervisor effort on the quality of answers. The upper panel shows the effect on high-quality ("good") answers and the lower panel shows the effect on low-quality ("bad") answers. Each observation represents an answer to one of 129,327 questions answered by 46 workers during 6,283 shifts.

Table 3-7: Effect of supervisor effort on additional measures of performance

(a) Probability that no volunteer leaves an answer

Outcome = No volunteer answers (mean = 0.486)				
	(1)	(2)	(3)	(4)
Supervisor effort (std)	0.0224*** (0.00255)	0.0194*** (0.00295)	0.0177*** (0.00249)	0.0109*** (0.00178)
Fixed effects				
Date and nightshift		x	x	x
Question category			x	x
Workers				x
N	129,327	129,327	129,327	129,327
R ²	0.00176	0.0166	0.0490	0.0534

Robust standard errors clustered by worker. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(b) Probability that a question expires

Outcome = A question expires (mean = 0.106)			
	(1)	(2)	(3)
Average supervisor effort (in shift)	-0.00890 (0.0408)	-0.0277 (0.0348)	-0.0284 (0.0348)
Fixed effects			
Date and nightshift		x	x
Question category			x
N	291,473	291,473	291,473
R ²	0.00000396	0.139	0.140

Robust standard errors clustered by shift. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: These tables report the effect of supervisor effort on the probability a volunteer answers a question and the probability that a question expires before being answered. For the upper panel, each observation represents one of 129,327 answers provided by workers. Of those questions, 48.6% were not answered by volunteers. For the lower panel, each observation represents one of 291,473 questions that either are answered, marked as invalid, or expired. We do not include worker fixed effects in this model since many questions expired before being picked up by a worker or after being completed by a worker and awaiting review by a supervisor. The variable *Average supervisor effort (in shift)* is an average of the on-duty supervisors' overall correction rate for answers. This variable is defined at the shift level and, therefore, we cluster standard errors at that same level. Relative to the sample presented in table 3-1, there are fewer observations because supervisor effort was not defined in shifts where reviews were not performed.

Table 3-8: Effect of supervisor effort on answer quality
(by worker experience)

	Outcome = "Good" answer (mean = 0.271)			Outcome = "Bad" answer (mean = 0.085)		
	(1)	(2)	(3)	(4)	(5)	(6)
Supervisor effort (std)	-0.00485* (0.00210)	-0.00467* (0.00225)	-0.00154 (0.00242)	-0.00459*** (0.00129)	-0.00481*** (0.00133)	-0.00334** (0.00109)
New worker × Supervisor effort (std)		-0.00100 (0.00508)	0.000935 (0.00453)		0.00136 (0.00197)	-0.000189 (0.00189)
New worker		0.00722 (0.0108)	0.00836 (0.0110)		-0.00333 (0.00301)	-0.00655 (0.00401)
Fixed effects						
Date and nightshift	x	x	x	x	x	x
Question category	x	x	x	x	x	x
Workers			x			x
N	129,327	129,327	129,327	129,327	129,327	129,327
R ²	0.0229	0.0229	0.0254	0.0142	0.0142	0.0156

Robust standard errors clustered by worker. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table reports the effect of supervisor effort on answer quality broken down by worker experience. Workers are considered "new" if, at the time they answered a question, they had previously answered 500 or fewer questions. Each observation represents an answer to one of 129,327 questions answered by 46 workers during 6,283 shifts. Columns 1-3 report the effect of supervision on good answers and columns 4-6 report the effect on bad answers. The estimates from columns 1 and 4 are identical to those in column 3 of tables 3-6a and 3-6b. A total of 16.4% of questions were answered by workers who were considered new workers at the time they provided answers.

Table 3-9: Decomposition of quality changes on corrected vs. uncorrected answers

	Was question corrected?		
	All	Corrected	Uncorrected
	answers	answers	answers
	(1)	(2)	(3)
Outcomes			
Good answer	-.0014	.0231***	-.0245***
Bad answer	-.0034**	.0077***	-.0111***
Size of quality change relative to fraction of corrections made			
Good answers	-.016	.272	-.288
Bad answers	-.040 ¹	.091	-.130
RF effect of Effort on Corrected	.0895		
Summary statistics			
	Outcome		
	Good Answer	Bad answer	
Uncorrected	0.2726	0.0847	
Corrected	0.2676	0.0856	
All answers	0.2711	0.0850	
1. Size of quality change relative to change in corrections = $\frac{\text{Reduction in badness}}{\text{RF effect of Effort on Corrected}} = \frac{-0.0034}{0.0895} = -.040$			
* p<0.05, ** p<0.01, *** p<0.001			

Notes: This table reports estimates for a decomposition of the total changes in quality for corrected vs. uncorrected answers (refer also to section 3.4.3). The *Outcomes* panel decomposes the total regression effect into the effect on corrected vs. uncorrected answers and the columns report the estimates from equations 3.3a, 3.3b, and 3.3c. Column 1 shows a regression of a quality outcome on supervisor effort. Column 2 shows a regression of the quality outcome times whether the answer was corrected and Column 3 shows a regression of the quality outcome times whether the answer was *not* corrected. The coefficients in columns 2 and 3 necessarily add to the coefficient in column 1. The second panel divides each coefficient in the *Outcomes* panel by the coefficient in the *RF effect of Effort on Corrected* row in order to show the magnitude of the change in quality relative to the change in corrections made. Footnote 1 provides an example that relates the reduction in badness to the fraction of corrections made. Finally, the third panel expresses the fraction of questions that were good and bad tabulated by whether they were corrected.

Table 3-10: IV estimates of the value of supervisor corrections

	No controls (1)	Date and nightshift (2)	Date, nightshift and categories (3)	All controls and Worker FE (4)
% "Good" answers	0.271			
% "Bad" answers	0.085			
Instrumental variable				
"Good" answer	-0.103** (0.0359)	-0.0816* (0.0355)	-0.0667* (0.0325)	-0.0156 (0.0251)
"Bad" answer	-0.0680*** (0.0185)	-0.0693** (0.0230)	-0.0631** (0.0211)	-0.0377** (0.0124)
Reduced form				
"Good answer"	-0.00844** (0.00267)	-0.00591** (0.00219)	-0.00485* (0.00210)	-0.00140 (0.00224)
"Bad answer"	-0.00556*** (0.00124)	-0.00502*** (0.00137)	-0.00459*** (0.00129)	-0.00338** (0.00103)
First stage				
Pr(correction) on effort	0.0818*** (0.00890)	0.0724*** (0.00895)	0.0727*** (0.00893)	0.0895*** (0.00712)
Fixed effects				
Date and nightshifts		X	X	X
Question categories			X	X
Workers				X
N	129,327	129,327	129,327	129,327

Robust standard errors clustered by worker. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: The row "First stage" shows a regression of the probability that a question was corrected on supervisor harshness, "Reduced form" shows a regression of the probability that an answer was "good" (i.e., received one or more thumbs-up) on supervisor harshness, and "Instrumental variable" show the IV estimate of the effect of a question correction on the probability that an answer was good. The first column doesn't include any controls, the second column includes fixed effects for each date and whether a shift was a night-shift, the third column adds additional fixed effects for the category of a question. Finally, column 4 includes all the controls as column 3, but adds fixed effects for each worker. In our sample, we exclude workers who also acted as supervisors and observe a total of 46 workers during 6,283 shifts.

Table 3-11: Variability by shift: Number of questions, answers, workers, and supervisors

	Mean	SD	10th	25th	50th	75th	90th
Questions per shift							
Period One	31.9	25.5	7	13	25	43	67
Period Two	62.3	40.6	18	31	54	82	124
Both periods	45.5	36.4	10	18	36	62	95
Answers per shift							
Period One	17.4	15.3	3	6	13	24	37
Period Two	36.2	26.7	9	16	29	50	71
Both periods	25.8	23.2	5	9	19	35	57
Workers per shift							
Period One	6.2	2.8	3	4	6	8	10
Period Two	12.3	7	5	7	10	17	24
Both periods	8.9	5.9	4	5	7	11	18
Supervisors per shift							
Period One	2.6	0.8	2	2	2	3	4
Period Two	3.1	1	2	2	3	4	4
Both periods	2.8	0.9	2	2	3	3	4

Notes: This table summarizes the variability in questions, answers, workers, and supervisors during each shift. Shifts are defined based on half-hour periods when questions arrive to be answered. Period one lasted 99 days and received 113,236 questions, of which 61,756 were answered, during 3,555 shifts. Period two lasted 81 days and received 179,911 questions, of which 104,490 were answered during 2,888 shifts. For each panel, the row *Both periods* is less informative for the quantile statistics because it combines periods one and two, which had very different levels of output. To best understand fluctuations during each period,

Table 3-12: Propensity of reviewers to make corrections

Reviewer classification	Reviewer Name	Company Role	# of actions (answers and reviews)	Fraction of actions that are reviews	"Supervisor Effort" (correction rate when reviewing)
	Alice	Team Leader	29,790	99%	21%
	Bob	Team Leader	17,308	97%	29%
	Chris	Worker	20,230	97%	27%
	Daniel	Team Leader	12,828	96%	31%
Heavy reviewer: More than 50% of actions were reviews	Emily	Team Leader	22,366	96%	45%
	Fred	Team Leader	967	96%	31%
	George	Team Leader	33,371	93%	16%
	Howard	Team Leader	5,721	90%	23%
	Ines	Worker	22,037	80%	23%
	Julian	Worker	21,692	79%	27%
	Kate	Worker	10,297	70%	45%
	Lucy	Worker	4,281	33%	20%
	Mark	Worker	7,147	20%	25%
	Neil	Worker	5,047	16%	61%
Light reviewer: Fewer than 50% of actions were reviews	Oscar	Worker	9,570	11%	32%
	Peter	Worker	7,967	7%	44%
	Mean		14,414		31%

Notes: This table lists all employees who made one or more reviews. "Company role" indicates whether the employee is a team leader or ordinary worker within the firm; although workers may perform a supervisory role on a particular project, team leaders are of a higher rank. The total number of actions is the number of questions answered or reviewed; and, reviewers are sorted by the fraction of actions that were reviews. "Supervisor Effort" is the fraction of questions each supervisor corrected during their reviews.

Table 3-13: Worker exposure to supervisors: Fraction of question-answers reviewed by their top-5 supervisors

Workers (greatest to fewest number of shifts)																								
Most frequent supervisors																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
1st	35%	18%	19%	23%	21%	26%	31%	28%	36%	33%	45%	49%	32%	30%	22%	25%	23%	50%	19%	19%	22%	38%	26%	
2nd	19%	17%	15%	19%	18%	16%	19%	23%	20%	22%	12%	16%	20%	14%	14%	18%	22%	29%	17%	15%	22%	15%	24%	
3rd	9%	16%	15%	16%	16%	16%	14%	14%	16%	11%	11%	12%	11%	13%	13%	13%	15%	5%	15%	14%	13%	12%	17%	
4th	8%	14%	15%	13%	15%	15%	9%	8%	7%	11%	11%	5%	10%	8%	11%	10%	13%	4%	10%	13%	9%	10%	11%	
5th	8%	14%	14%	8%	11%	12%	7%	7%	6%	5%	7%	5%	8%	7%	11%	9%	8%	4%	10%	9%	8%	7%	7%	
Top Five	79%	79%	78%	78%	81%	85%	80%	80%	85%	82%	85%	89%	82%	72%	71%	75%	81%	92%	72%	70%	75%	81%	84%	
All others	21%	21%	22%	22%	19%	15%	20%	20%	15%	18%	15%	11%	18%	28%	29%	25%	19%	8%	28%	30%	25%	19%	16%	
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46																								
1st	45%	21%	19%	30%	22%	28%	21%	28%	43%	20%	29%	26%	31%	20%	63%	20%	48%	25%	45%	27%	34%	57%	49%	
2nd	26%	18%	18%	18%	18%	14%	18%	21%	32%	16%	22%	17%	16%	16%	27%	16%	16%	21%	14%	24%	13%	28%	34%	
3rd	12%	13%	13%	10%	14%	10%	18%	15%	6%	15%	10%	13%	10%	13%	3%	15%	12%	17%	7%	23%	12%	15%	17%	
4th	5%	12%	10%	9%	12%	10%	13%	10%	4%	14%	9%	13%	10%	12%	2%	10%	9%	15%	6%	8%	12%	0%	0%	
5th	3%	9%	8%	9%	10%	10%	11%	6%	3%	10%	6%	10%	9%	11%	2%	9%	6%	8%	6%	8%	11%	0%	0%	
Top Five	92%	73%	69%	76%	75%	71%	81%	79%	89%	74%	76%	78%	75%	71%	96%	70%	92%	85%	79%	90%	82%	100%	100%	
All others	8%	27%	31%	24%	25%	29%	19%	21%	11%	26%	24%	22%	25%	29%	4%	30%	8%	15%	21%	10%	18%	0%	0%	

Notes: This table illustrates shows the fraction of shifts that each worker spent under different supervisors. Each column represents a worker and the worker's number indicates how she ranks compared to other workers in terms of the number of shifts worked (1 = most shifts, 46 = fewest shifts). The first five rows represents the fraction of a worker's shifts when she was on-duty with each of her top five supervisors and the row *Top Five* sums up the proportion of shifts with top five supervisors. The row *All others* represents the fraction of shifts with all other supervisors.

Bibliography

- Alatas, Vivi, Banerjee Abhijit, Rema Hanna, Benjamin A Olken, Ririn Purnamasari, and Matthew Wai-poi**, “Ordeal mechanisms in targeting: theory and evidence from a field experiment in indonesia,” *Working Paper*, 2012.
- Angrist, Joshua D**, “Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors,” *Journal of Business & Economic Statistics*, 2001, 19 (1), 2–28.
- Ariely, Dan, Emir Kamenica, and Drazen Prelec**, “Man’s Search for Meaning: The Case of Legos,” *Journal of Economic Behavior & Organization*, 2008, 67 (3-4), 671 – 677.
- Autor, David H**, “Wiring the labor market,” *Journal of Economic Perspectives*, 2001, 15 (1), 25–40.
- Avery, Christopher and Jonathan D Levin**, “Early Admissions at Selective Colleges,” *American Economic Review*, 2010, 100.
- Barankay, I.**, “Rankings and social tournaments: Evidence from a field experiment,” *University of Pennsylvania mimeo*, 2010.
- Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz**, “Evaluating Online Labor Markets for Experimental Research: Amazon.com’s Mechanical Turk,” *Political Analysis*, 2012, 20, 351–368.
- Besley, Timothy and Stephen Coate**, “Workfare versus Welfare : Incentive Arguments for Work Requirements in Poverty-Alleviation Programs,” *American Economic Review*, 1992, 82 (1), 249–261.
- Cameron, A Colin and Pravin K Trivedi**, *Microeconometrics: methods and applications*, Cambridge university press, 2005.
- Coles, Peter, Alexey Kushnir, and Muriel Niederle**, “Preference signaling in matching markets,” Technical Report, National Bureau of Economic Research 2010.

- , J Cawley, P B Levine, M Niederle, Alvin E Roth, and J J Siegfried, “The job market for new economists: A market design perspective,” *The Journal of Economic Perspectives*, 2010, 24 (4), 187–206.
- Crosen, R. and U. Gneezy**, “Gender differences in preferences,” *Journal of Economic Literature*, 2009, 47 (2), 448–474.
- Doyle, Joseph J**, “Child Protection and Child Outcomes: Measuring the Effects of Foster Care,” *American Economic Review*, 2007, 97 (5), 1583–1610.
- Eriksson, K. and B. Simpson**, “Emotional reactions to losing explain gender differences in entering a risky lottery,” *Judgment and Decision Making*, 2010, 5 (3), 159–163.
- Gneezy, Uri and John A. List**, “Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments,” *Econometrica*, 2006, 74 (5), 1365–1384.
- Harrison, Glenn W. and John A. List**, “Field Experiments,” *Journal of Economic Literature*, 2004, 42 (4), 1009 – 1055.
- Henrich, J., S.J. Heine, A. Norenzayan et al.**, “The weirdest people in the world,” *Behavioral and Brain Sciences*, 2010, 33 (2-3), 61–83.
- Holmes, Susan and Adam Kapelner**, “DistributeEyes,” *Stanford University Manuscript*, 2010.
- Horton, John J. and Lydia Chilton**, “The Labor Economics of Paid Crowdsourcing,” *Proceedings of the ACM Conference on Electronic Commerce, Forthcoming*, 2010.
- , **David G. Rand, and Richard J. Zeckhauser**, “The Online Laboratory: Conducting Experiments in a Real Labor Market,” *Experimental Economics*, 2011, 14, 399–425.
- Ipeirotis, Panos**, “Demographics of Mechanical Turk,” CeDER working paper CeDER-10-01, New York University, Stern School of Business. March 2010.
- Kling, Jeffrey**, “Incarceration length, employment, and earnings,” *American Economic Review*, 2006, 75 (3), 424–440.
- Kuhn, Peter and Hani Mansour**, “Is Internet Job Search Still Ineffective?,” *The Economic Journal*, 2013.
- **and Mikal Skuterud**, “Internet Job Search and Unemployment Durations,” *The American Economic Review*, 2004, 94 (1), 218–232.

- Lazear, Edward P.**, "Performance Pay and Productivity," *American Economic Review*, 2000, 90 (5), 1346–1361.
- Lazear, E.P., K.L. Shaw, and C.T. Stanton**, "The Value of Bosses," *NBER Working Paper*, 2011.
- Lee, Soohyung, Muriel Niederle, Hye-Rim Kim, and Woo-Keum Kim**, "Propose with a rose? signaling in internet dating markets," Technical Report, National Bureau of Economic Research 2011.
- Levitt, S.D. and J.A. List**, "Field experiments in economics: the past, the present, and the future," *European Economic Review*, 2009, 53 (1), 1–18.
- Levitt, Steven D. and John A. List**, "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives*, 2007, 21 (2), 153–174.
- Maggio, Marco Di and Marshall Van Alstyne**, "Information Sharing and Individual Performance : Evidence from a Japanese Bank," *Working Paper*, 2011.
- Mas, Alexandre and Enrico Moretti**, "Peers at Work," *American Economic Review*, 2009, 99 (1), 112–145.
- Nagin, Daniel S, James B Rebitzer, Seth Sanders, and Lowell J Taylor**, "Monitoring , Motivation , and Management : The Determinants of Behavior in a Field Experiment Opportunistic," *American Economic Review*, 2002, 92 (4), 850–873.
- Nichols, Albert L and Richard J Zeckhauser**, "Targeting transfers through restrictions on recipients," *The American Economic Review*, 1982, pp. 372–377.
- Paolacci, G., J. Chandler, and P.G. Ipeirotis**, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, 2010, 5 (5), 411–419.
- Preston, Anne E.**, "The Nonprofit Worker in a For-Profit World," *Journal of Labor Economics*, 1989, 7 (4), 438–463.
- Rosen, S.**, "The theory of equalizing differences," *Handbook of labor economics*, 1986, 1, 641–692.
- Rosso, Brent D., Kathryn H. Dekas, and Amy Wrzesniewski**, "On the meaning of work: A theoretical integration and review," *Research in Organizational Behavior*, 2010, 30, 91–127.

- Sprouse, J.**, "A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory," *Behavior Research Methods*, 2011, 43 (1), 155–167.
- Stern, S.**, "Do Scientists Pay to Be Scientists?," *Management Science*, 2004, 50 (6), 835–853.
- Stevenson, Betsey**, "The Internet and job search," Technical Report, National Bureau of Economic Research 2008.
- Tuck, Laura and Kathy Lindert**, "From Universal Food Subsidies to a Self-Targeted Program: A Case Study in Tunisian Reform," *World Bank Discussion Paper No. 351*, 1996.
- Varian, Hal R.**, "Computer Mediated Transactions (2010 AEA Keynote)," *American Economic Review*, 2010, 100 (2).
- Wrzesniewski, Amy, Clark Mccauley, Paul Rozin, and Barry Schwartz**, "Jobs, Careers, and Callings: People's Relations to Their Work Amy Wrzesniewski," *Journal of Research in Personality*, 1997, 33 (31), 21–33.