John J. Horton: Research Papers

Published & Forthcoming Papers in Economics

"The Online Laboratory: Conducting Experiments in a Real Labor Market" (with David Rand and Richard Zeckhauser) Experimental Economics, 14:3 (2011), 399-425.

"The Condition of the Turking Class: Are Online Employers Fair and Honest?" Economics Letters, 111:1 (April 2011), 10-12.

Published & Forthcoming Papers in Computer Science

"Labor Allocation in Paid Crowdsourcing: Experimental Evidence on Positioning, Nudges and Prices" (with Dana Chandler) Proceedings of the 25th Conference on Artificial Intelligence (AAAI), Human Computation Workshop (HCOMP), August 2011

"Designing Incentives for Inexpert Human Raters" (with Aaron Shaw and Daniel Chen) Proceedings of the ACM Conference of Computer Supported Cooperative Work (ACM-CSCW), Best Paper Nominee, March 2011

"Online Labor Markets" Proceedings of the 6th Workshop on Internet and Network Economics (WINE), December 2010

"Algorithmic Wage Negotiations: Applications to Paid Crowdsourcing" (with Richard Zeckhauser) Proceedings of CrowdConf, 2010

"Task Search in a Human Computation Market" (with Lydia Chilton, Rob Miller and Shiri Azenkot) Proceedings of the ACM Conference on Knowledge Discovery and Data Mining/Human Computation (ACM-KDD/HCOMP), 2010

"The Labor Economics of Paid Crowdsourcing" (with Lydia Chilton) Proceedings of the 11th ACM Conference on Electronic Commerce (ACM-EC), 2010

Working Papers (Not including Job Market Paper)

"Employer Expectations, Peer Effects and Productivity: Evidence from a Series of Field Experiments" 2010. Status: Draft available.

"The Wages of Pay Cuts: Evidence from a Field Experiment" (with Daniel Chen) 2012. Status: Draft available.

"Procurement, Incentives and Bargaining Friction: Evidence from Government Contracts" 2009. Status: R&R at The Journal of Law & Economics

Resting Papers

"The Dot Guessing Game: A Fruit Fly For Human Computation" 2010. Status: Draft available.

Position Papers and Other Non-refereed publications

"The Need for Standardization in Crowdsourcing" (with Panagiotis Ipeirotis) Presented at the Crowdsourcing Workshop, ACM-CHI 2011 (Vancouver, BC), May 2011

"Heads in the Cloud: Challenges and Opportunities in Human Computation" (with Robert C. Miller, Greg Little, Michael Bernstein, Jeffrey P. Bigham, Lydia B. Chilton, Max Goldman and Rajeev Nayak) XRDS: Crossroads, the ACM Magazine for Students, December 2010

Exp Econ (2011) 14:399–425
DOI 10.1007/s10683-011-9273-9

The online laboratory: conducting experiments in a real labor market

John J. Horton · David G. Rand · Richard J. Zeckhauser

Received: 28 June 2010 / Accepted: 11 January 2011 / Published online: 20 February 2011

Abstract Online labor markets have great potential as platforms for conducting experiments. They provide immediate access to a large and diverse subject pool, and allow researchers to control the experimental context. Online experiments, we show, can be just as valid—both internally and externally—as laboratory and field experiments, while often requiring far less money and time to design and conduct. To demonstrate their value, we use an online labor market to replicate three classic experiments. The first finds quantitative agreement between levels of cooperation in a prisoner's dilemma played online and in the physical laboratory. The second shows—consistent with behavior in the traditional laboratory—that online subjects respond to priming by altering their choices. The third demonstrates that when an identical decision is framed differently, individuals reverse their choice, thus replicating a famed Tversky-Kahneman result. Then we conduct a field experiment showing that workers have upward-sloping labor supply curves. Finally, we analyze the challenges to online experiments, proposing methods to cope with the unique threats to validity in an online setting, and examining the conceptual issues surrounding the external validity of online results. We conclude by presenting our views on the potential role that online experiments can play within the social sciences, and then recommend software development priorities and best practices.

Thanks to Alex Breinin and Xiaoqi Zhu for excellent research assistance. Thanks to Samuel Arbesman, Dana Chandler, Anna Dreber, Rezwan Haque, Justin Keenan, Robin Yerkes Horton, Stephanie Hurder and Michael Manapat for helpful comments, as well as to participants in the Online Experimentation Workshop hosted by Harvard's Berkman Center for Internet and Society. Thanks to Anna Dreber, Elizabeth Paci and Yochai Benkler for assistance running the physical laboratory replication study, and to Sarah Hirschfeld-Sussman and Mark Edington for their help with surveying the Harvard Decision Science Laboratory subject pool. This research has been supported by the NSF-IGERT program "Multidisciplinary Program in Inequality and Social Policy" at Harvard University (Grant No. 0333403), and DGR gratefully acknowledges financial support from the John Templeton Foundation's Foundational Questions in Evolutionary Biology Prize Fellowship.

J.J. Horton · D.G. Rand · R.J. Zeckhauser (✉)
Harvard University, Cambridge, USA
e-mail: richard_zeckhauser@harvard.edu

Keywords Experimentation · Online labor markets · Prisoner's dilemma · Field experiment · Internet

JEL Classification J2 · C93 · C91 · C92 · C70

1 Introduction

Some of the first experiments in economics were conducted in the late 1940s to test predictions from the emerging field of game theory. While the research questions were complex, the tools were simple; paper, pencils, blackboards and marbles were sufficient instruments to present stimuli and capture subjects' choices and actions.[1]

By the early 1990s, researchers had developed tools for conducting experiments over local computer networks, with subjects receiving stimuli and making decisions via computer terminal (Fischbacher 2007). This development made it easier to carry out game play and collect data. However, the advantages of computer-mediation were not merely logistical; experimenters also gained greater control over the flow of information and thereby reduced the relevance of potentially confounding factors. For these reasons, computer-mediation quickly became the primary means for conducting laboratory experiments.

Today, human subjects are still brought into physical laboratories despite the fact that many, if not most, computer-mediated experiments can easily be conducted over the Internet. Online participation would spare experimenters the expenses of physically aggregating subjects and compensating them for travel. This in turn would allow for larger and longer games with potentially more diverse subjects. Social scientists recognized these advantages more than a decade ago. In 1997, the National Science Foundation sponsored a workshop called NetLab to investigate the potential of online experimentation (Bainbridge 2007). That workshop's report identified the major advantages of online experimentation and optimistically concluded:

> If the nascent laboratory experimental approach is encouraged and is coupled with new technological innovations, then the SBE [social, behavioral, and economic sciences] disciplines will be primed for major scientific advances.

The "if" in that conclusion hinged on some rather mundane obstacles: funding, particularly for software development, and technical training. Thirteen years have passed; and yet, despite an explosion in the size, usage, and capabilities of the Internet, online experiments are still relatively rare, particularly in economics. During the same period, both field and traditional laboratory experiments have become far more common (Levitt and List 2009). We believe that the practical problems of (1) recruiting subjects and paying them securely and (2) assuring internal validity—and not those of funding or training constraints—have limited the development of online experimentation.

In this paper, we argue that a recent development effectively and efficiently addresses both the recruitment/payment problem and the internal validity problem. This

[1] For an historical perspective on early experimentation in economics, see Kagel et al. (1995).

development is the emergence of online labor markets. In these markets, workers from around the world perform tasks amenable to remote completion, such as data entry, computer programming, graphic design and clerical work (Frei 2009). These markets, although designed for other purposes, make it possible to recruit large numbers of subjects who are ready and able to participate in experiments.

These recruited subjects have the attractive properties of being diverse and not experiment-savvy. But their key characteristic is that they will participate in the experiment within the context of an online labor market. This is critical, because the creators of online labor markets—for their own, non-experimental purposes—have built their platforms in a way that grants experimenters the control needed for valid causal inference. In particular, the creators of these markets have made it easy to make individual-specific payments, screen out users who do not have valid accounts with the market and prevent workers/subjects from communicating with each other.

Despite the benefits they offer, online experiments raise issues not frequently encountered in either the laboratory or the field. Just as television shows are not filmed plays, online experiments are not simply laboratory experiments conducted online. This paper identifies the major differences and pays close attention to the unique challenges of online experimentation. Despite the caveats and potential pitfalls, the value of online experiments is demonstrated by our replications; we quickly, cheaply and easily reproduce a handful of experimental results known to have external validity. Given that online experiments work, at least for the cases we tried, the logical next question is why. Much of the material that follows seeks to answer that question.

In Sect. 2 we provide information on online labor markets and discuss in broad terms how these markets allow researchers to overcome the classic challenges to causal inference. We also discuss the strengths and inherent limitations of the online laboratory. In Sect. 3, we successfully reproduce the qualitative characteristics of a series of classic experimental results, as well as discuss other examples of research making use of these markets. These confirmations support our basic argument, but challenges remain. In Sect. 4, we address the primary specific challenges of conducting experiments online, and provide tentative solutions to these challenges. In Sect. 5 we discuss the external validity of online experimental results. In Sect. 6, we analyze different experimental designs that can be used online. In addition to creating exciting opportunities for research, online experiments also pose particular ethical challenges. They are the subject of Sect. 7. We conclude in Sect. 8 with our thoughts on the future of the online laboratory.

2 Overview of experimentation in online labor markets

At present, the most useful online labor markets, from an experimentation standpoint, are "all-purpose" labor markets where buyers contract with individual sellers (Horton 2010). Some of the larger markets in this category include oDesk, Freelancer, Elance,

Guru and Amazon's Mechanical Turk (MTurk).[2] Each of these sites is potentially amenable to experimentation, but MTurk currently offers the best venue due to its robust application programming interface (API) and its simple yet flexible pricing structure.

Online experiments are quite easy to run: an advertisement is placed for the experiment via the same framework used to advertise real jobs on the market. This advertisement offers a general description of the experiment that truthfully encompasses all the experimental groups. As subjects accept this "job," they are assigned by the experimenter (usually with the aid of a random number generator) to an experimental group. Each group encounters a different interface according to their group assignment. For example, interfaces might differ on instructions, payment schedules or visual stimuli. The interface is usually a stand-alone website that gives the subjects instructions, records their choices, provides them with information as the game progresses and determines their payoffs. After subjects complete whatever task is asked of them (e.g., perform work or make choices), they "submit" the task and are eventually paid, just like they would for any other work performed in the market.

There are now several papers that serve as "how to" guides for running behavioral experiments specifically on Mechanical Turk. Paolacci et al. (2010) and Mason and Watts (2010) both focus on the practical challenges of running experiments on MTurk, and serve as excellent resources for getting started. One practical advantage of MTurk is that it supports experiments that range from simple surveys made with off-the-shelf or web-based software to custom-built, elaborate interfaces with designs limited only by time and resources.

2.1 The advantages of recruiting from labor markets

Subjects recruited from MTurk, or from any online labor market, potentially provide diverse samples of both high- and low-skilled individuals from a wide range of countries. One potentially useful dimension of subject diversity is inexperience with economic games, though the magnitude of this advantage is likely to depend on the research question. Further, by using subjects from less-developed countries, experimenters can create relatively high-stakes games for far less money than would be needed if using subjects from developed countries.

Depending on a researcher's institution and research question, experimenters might not be required to tell subjects hired from online labor markets that they are participating in an experiment. For experiments that use classic economic games, subjects might guess they are in a study of some kind; but for real-effort, market-appropriate tasks, workers are unlikely to suspect that their "employer" is a researcher. This advantage partially answers one of the sharpest critiques of the experimental method in economics, namely the inherent artificiality created by subjects knowing they are in an experiment. Subjects recruited from online labor markets are already making consequential economic decisions and they are likely to view any

[2] There are other online labor markets, structured more like tournaments or prize-based contests, that are less relevant for experimental purposes.

task or game using an economic frame of mind. Even in non-economic scenarios, lack of subject awareness is useful, as their uninformed state rules out experimenter effects and John Henry effects.[3]

When subjects are aware they are in an experiment, they might try to learn about the conditions of experimental groups. Subjects in a less desirable treatment might be upset by their bad luck, which might affect their behaviors. Cook and Campbell (1979) call this "demoralization." Furthermore, even if subjects remain unaware of an experiment and of the nature of the treatment, agents of the experimenter might affect outcomes through their own initiative, such as by compensatory equalization (i.e., intervening to make the outcomes of the different groups more similar and hence "fairer"). In online experiments in which subjects have no knowledge of the treatments received by others, the threat of demoralization is minimal, and since carrying out online experiments generally requires no human agent, unauthorized interventions like compensatory equalization are unlikely.

2.2 Obtaining control and promoting trust

Even though online labor markets provide a pool of would-be subjects with some desirable characteristics, having subjects alone is not sufficient for the conduct of an experiment. Experimenters need to be able to uniquely identify these subjects, convey instructions, collect responses and make payments, while being confident that their intended actions are actually being implemented properly. Fortunately, many of the primary concerns of would-be experimenters mirror the concerns of the creators and customers of online labor markets. Employers worry that workers with multiple accounts might place phony bids or manipulate the reputation system by leaving phony feedback. Similarly, experimenters worry that a subject with multiple accounts might participate in an experiment multiple times. The creators of online labor markets do not want workers to communicate with each other, as that could lead to collusion. Experimenters also worry about workers discussing the details of experiments with each other and possibly colluding. Finally, both employers and experimenters need ways to pay individuals precise amounts of money as rewards for their actions and decisions.

It is now easy to hire and pay workers within the context of online labor markets, yet still quite difficult to do the same online, but outside of these markets. The problem is not technological. The type and quality of communication—email, instant messenger services, voice-over-IP—do not depend on whether the buyer and seller are working inside or outside the online market, and banks have been transferring funds electronically for decades. The problem is that it is difficult to create trust among strangers. Trust is an issue not only for would-be trading partners, but also for would-be experimenters.

The validity of economics experiments depends heavily upon trust, particularly subjects' trust that the promulgated rules will be followed and that all stated facts

[3] Experimenter effects are created when subjects try to produce the effect they believe the experimenters expect; "John Henry" effects are created when subjects exert great effort because they treat the experiment like a competitive contest.

about payment, the identities of other subjects, etc., are true. This need for trust provides a good reason to embed experiments in online labor markets, because the creators of these markets have already taken a number of steps to foster trust of employers. The issue of trust is so critical that we investigate it empirically (in Sect. 5.4) via a survey of subjects recruited from both MTurk and a subject pool used for traditional laboratory experiments.

All major online labor markets use reputation systems to create lasting, publicly-available reputations—reputations that are sacrificed if either buyers or workers behave unfairly (Resnick et al. 2000). The market creators proactively screen out undesired participants by taking steps, such as requiring a bank account or valid credit card, before either buyer or seller is allowed to join. With persons who have been accepted, the market creators actively manage memberships and suspend bad actors, creating a form of virtuous selection not found in traditional markets.

One kind of bad actor is the non-human, automated script that fraudulently performs "work." To combat this potential problem, all sites require would-be members to pass a CAPTCHA, or "completely automated public Turing test to tell computers and humans apart" (von Ahn et al. 2003). At least on MTurk, there is some danger of malicious users writing scripts that automatically accept and complete "Human Intelligence Tasks," or HITs. However, these attempts are trivially easy to detect for anything more complicated than a single yes/no question. Furthermore, asking comprehension questions regarding the details of the experimental instructions, as well as recording the total time taken to complete the HIT, allows experiments to distinguish automated responders from actual subjects. In our experience, jobs that allow workers to only complete one unit of work (which is almost always the case with experiments) do not attract the attention of scammers writing scripts (because would-be scammers cannot amortize script-writing costs over a larger volume of work). With proper precautions, it is unlikely that computers would show up as subjects, or that any workers/subjects would believe they were playing against a computer.

While the "Turing test" form of trust is important, the mundane but perhaps more critical requirement is that workers/subjects trust that buyers/experimenters will actually follow the rules that they propose. To encourage this form of trust, many online labor markets require buyers to place funds in escrow, which prevents buyers from opportunistically refusing to pay after taking delivery of the worker's output (which is often an easy-to-steal informational good). In many online markets, there is some form of dispute arbitration, which encourages the belief that all parties are operating in the shadows of an institution that could hold them accountable for their actions, further promoting trust. Perhaps unsurprisingly, survey evidence suggests that workers in MTurk believe that their online bosses are as fair as employers in their home countries (Horton 2011).

2.3 Limitations of online experiments

Online experiments, like any experimental method, have limitations, even when conducted within online labor markets. One of the most obvious is that only some types of experiments can be run. Surveys and one-shot "pen and pencil"-style economic games are extremely straightforward and therefore amenable. Repeated games are

also possible given the proper software tools (see Suri and Watts 2011 as an example). However, designs that require the physical presence of participants are clearly impossible. For example, recording physiological responses like eye movement, the galvanic skin response or blood flow to the brain cannot be done online; neither can interventions which involve physically manipulating the subjects such as having them touch hot versus cold objects, nor manipulating the subjects' environment such as changing lighting levels or playing loud music. Face to face communication is also challenging, although potentially surmountable to some extent given the widespread adoption of webcams and other video-chat technologies.

A further limitation is the difficulty of creating common knowledge among participants. In the traditional lab, it is possible to read the instructions aloud, such that participants know that everyone has received the same instructions. Online, the best that can be done is to inform subjects that all participants receive the same instructions, but this cannot be verified by the subjects. On the other hand, it is possible to build engaging instructional materials and to conduct in-game tests of comprehension before allowing subjects to continue in an experiment, thus making it more likely that all subjects do in fact possess common knowledge.

At present there is also no easy way to answer questions that subjects may have about the instructions, though in principle experimenters could communicate with subjects through email, VoIP or chat. However, this kind of interaction is more complicated and burdensome than the immediate feedback that can be given in a laboratory. This difficulty puts some limits on the complexity of experiments that can be easily run on MTurk. One way to deal with this issue is to include comprehension questions that verify subjects' understanding of the experiment, with correct answers being a prerequisite for participation. Although many subjects might fail for complicated tasks, experiments can take advantage of the large number of potential participants on MTurk to continue recruiting until enough comprehending subjects have accumulated.

One potentially serious limitation to online experimentation is uncertainty about the precise identity of the experimental subjects. We discuss this problem at length in Sect. 4, but we admit that 100% confidence that the problem is ruled out is unrealistic. Even if we are certain that each worker has one and only one account, it is possible that multiple workers share the same account. Thus it is possible that different people will complete different sections of a single study, or that several people will work together to complete a single set of decisions. This raises various potential challenges in terms of consistency of responses across sections, judging the effort invested by participants, and group versus individual decision-making (Kocher and Sutter 2005). A partial solution is provided by online labor markets that map more closely to traditional markets (like oDesk and Elance) and that provide more elaborate tools for verifying worker identities.

3 Experiments in the online laboratory

This section discusses research conducted in the online laboratory, both by ourselves and others. We conducted three laboratory experiments for this paper, one a direct

quantitative replication of an experiment we ran in the physical laboratory, and two qualitative replications of experiments with well-known and widely reproduced results.

These studies provide evidence that subjects on MTurk behave similarly to subjects in physical laboratories. These successful replications suggest that online experiments can be an appropriate tool for exploring human behavior, and merit a place in the experimentalist's toolkit alongside traditional offline methods, at least for certain research questions.

Our first experiment had subjects play a traditional one-shot prisoner's dilemma game. We conducted the experiment both on MTurk and in the physical laboratory. The experimental design was the same, except that the MTurk payoffs were 10 times smaller than the payoffs in the physical lab. We found no significant difference in the level of cooperation between the two settings, providing a quantitative replication of physical lab behavior using lower stakes on MTurk. In both settings, a substantial fraction of subjects displayed other-regarding preferences.

Our second experiment had subjects play the same prisoner's dilemma game, after having been randomly assigned to read different "priming" passages of religious or non-religious text. Here we demonstrated the well-established fact that stimuli unrelated to monetary payoffs can nonetheless affect subjects' decisions. In both the second and third experiments, subjects earned individualized payments based on their choices and the choices of other workers with whom they were randomly matched retroactively.

Our third experiment replicated a famed result in framing shown by Tversky and Kahneman (1981). In accordance with numerous duplications in the laboratory, we found that individuals are risk-averse in the domain of gains, and risk-seeking in the domain of losses. Subjects were paid a fixed rate for participating.

Beyond our laboratory experiments, we conducted a *natural field experiment* in the sense of the taxonomy proposed by Harrison and List (2004). It looked at the labor supply response to manipulations in the offered wage. This experiment placed us in the role of the employer. This experimenter-as-employer research design is perhaps the most exciting development made possible by online labor markets. We recruited subjects to perform a simple transcription of a paragraph-sized piece of text. After performing this initial task, subjects were offered the opportunity to perform an additional transcription task in exchange for a randomly determined wage. As expected, we found that workers' labor supply curves slope upward.

3.1 Existing research

Several studies using online subject pools have appeared recently, with computer scientists leading the way. They all used MTurk, primarily as a way to conduct user-studies and collect data suitable for the training of machine learning algorithms (Sheng et al. 2008; Kittur et al. 2008; Sorokin and Forsyth 2008). In a paper that bridges computer science and economics, Mason and Watts (2009) showed that, although quality is not affected by price, output declines when wages are lowered.

Among the several economics papers that used online labor markets, Chen and Horton (2010) measured the way MTurk workers respond to wage cuts. They found

that unexplained wage cuts decrease output, but that when the cuts were justified to workers, the former levels of output were maintained.[4] In a separate paper using MTurk, Horton and Chilton (2010) explored whether a simple rational model can explain worker output. While they found strong evidence that at least some workers are price-sensitive, they also found that a non-trivial fraction are target earners, that is, people who work to achieve certain income targets rather than responding solely to the offered wage. In a third study, Suri and Watts (2011) had MTurk subjects play the same repeated public goods game run in the physical laboratory by Fehr et al. (2000). Their MTurk subjects quantitatively replicated the experimental findings from the physical lab, using an order of magnitude lower payoffs. In a natural field experiment conducted on MTurk, Chandler and Kapelner (2010) subtly manipulated the meaning of the task and measured whether that change affected uptake and work quality, both overall and conditional upon a worker's home country. Their work illustrates the kinds of experiments that would be very difficult and costly to conduct in offline settings.

In addition to conventional academic papers, a number of researchers are conducting experiments on MTurk and posting results on their blogs. Gabriele Paolacci at the University of Venice writes a blog called "Experimental Turk" which focuses on reproducing results from experimental psychology.[5]

While MTurk is to date the mostly commonly used online labor market, others are emerging. For example, Pallais (2010) conducted a field experiment on the online labor market oDesk, in which she invited a large number of workers to complete a data-entry task. She found that obtaining a first job and receiving a feedback score helped them obtain future work in the market.

3.2 Quantitative replication: social preferences

A central theme in experimental economics is the existence of social (or "other-regarding") preferences (Andreoni 1990; Fehr and Schmidt 1999). Countless laboratory experiments have demonstrated that many people's behaviors are inconsistent with caring only about their own monetary payoffs. (For a review, see Camerer 2003.) Here we quantitatively replicated the existence and extent of other-regarding preferences in the online laboratory using MTurk.

To compare pro-social behavior on MTurk to that which is observed in the physical laboratory (hereafter often referred to as 'offline'), we used the prisoner's dilemma ("PD"), the canonical game for studying altruistic cooperation (Axelrod and Hamilton 1981). We recruited 155 subjects on MTurk and 30 subjects at Harvard University, using the same neutrally-framed instructions, incentive-compatible design and ex-post matching procedure. To be commensurate with standard wages on MTurk,

[4] There are a number of papers that have used the Internet as a test bed for field experimentation, primarily as a way to study auctions (Resnick et al. 2006; Lucking-Reiley 2000).

[5] Although blogs are certainly not equivalent to peer-reviewed journals, they do allow academics to quickly communicate results and receive feedback. For example, Rob Miller and Greg Little at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) host a blog called "Deneme" that reports the results of experiments using TurKit—a Java library developed by Little and others to perform iterative, complex tasks on MTurk (Little et al. 2009).

payoffs were an order of magnitude smaller on MTurk compared to the offline lab. MTurk participants received a \$0.50 "show-up fee," while offline subjects received a \$5 show-up fee. Each subject was informed that he or she had been randomly assigned to interact with another participant.[6] They were further informed that both players would have a choice between two options, A or B, (where A represents cooperation and B represents defection) with the following payoff structure:

$$\text{MTurk: } \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} \$0.70, \$0.70 & \$0, \$1.00 \\ \$1.00, \$0 & \$0.30, \$0.30 \end{pmatrix} \end{array} \quad \text{Physical lab: } \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} \$7, \$7 & \$0, \$10 \\ \$10, \$0 & \$3, \$3 \end{pmatrix} \end{array} \tag{1}$$

Note that, regardless of the action of one's partner, choosing B maximizes one's payoff. MTurk workers were additionally given five comprehension questions regarding the payoff structure, allowing us to compare subjects who were paying close attention with those who were not.[7]

In a one-shot PD, rational self-interested players should always select B. Consistent with a wealth of previous laboratory studies (Camerer 2003), however, a substantial fraction of our subjects chose A, both offline (37%) and on MTurk (47%). The difference between offline and online cooperation was not statistically significant (χ^2 test, $p = 0.294$), although this may have been in part due to the relatively small offline sample size ($N = 30$). However, if we restrict our attention to the $N = 74$ MTurk subjects who correctly answered all five comprehension questions, we find that 39% chose A, giving close quantitative agreement with the 37% of cooperating physical laboratory subjects (χ^2 test, $p = 0.811$). See Fig. 1. These results demonstrate the ability of MTurk to quantitatively reproduce behavior from the physical laboratory, and also emphasize the importance of using payoff comprehension questions in the context of MTurk.

3.3 Qualitative replication: priming

Priming is a common tool in the behavioral sciences. In priming studies, stimuli unrelated to the decision task (and which do not affect the monetary outcomes) can nonetheless significantly alter subjects' behaviors. Priming has attracted a great deal of attention in psychology, and, more recently, in experimental economics (Benjamin et al. 2010a). In our second experiment, we demonstrated the power of priming effects on MTurk.

[6] MTurk subjects were matched exclusively with MTurk subjects, and offline subjects matched exclusively with offline subjects.

[7] The comprehension questions were: (1) Which earns you more money: [You pick A, You pick B]. (2) Which earns the other person more money: [You pick A, You pick B]. (3) Which earns you more money: [Other person picks A, Other person picks B]. (4) Which earns the other person more money: [Other person picks A, Other person pick B]. (5) If you pick B and the other picks A, what bonus will you receive?

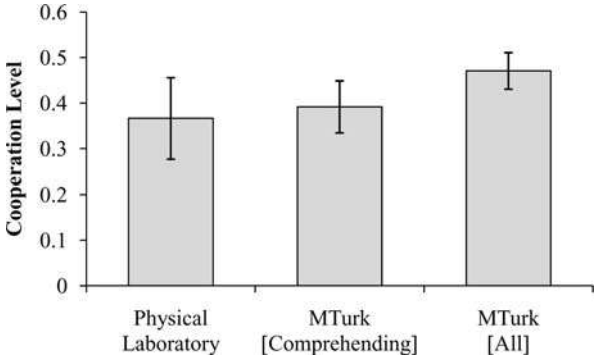


Fig. 1 Cooperation level in a one-shot prisoner's dilemma is similar amongst physical laboratory subjects, MTurk workers who correctly answered five payoff comprehension questions, and all MTurk workers. *Error bars* indicate standard error of the mean

To do so, we recruited 169 subjects to play a PD game. In addition to a $0.20 show-up fee, subjects were further informed of the following payoff structure:

$$\begin{array}{cc} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} \$1.20, \$1.20 & \$0.40, \$1.60 \\ \$1.60, \$0.40 & \$0.80, \$0.80 \end{pmatrix} \end{array} \tag{2}$$

As in the previous PD, A represents cooperation and B represents defection. Subjects were randomly assigned to either the religious prime group ($N = 87$) or a neutral prime group ($N = 82$). The religious prime group read a Christian religious passage about the importance of charity (Mark 10:17–23) before playing the PD. The neutral prime group instead read a passage of equal length describing three species of fish before playing the PD. Following the PD, each subject completed a demographic questionnaire reporting age, gender, country of residence and religious affiliation. The subjects also indicated whether they had ever had an experience which convinced them of the existence of God (here called "believers"). Based on previous results using implicit primes with a non-student subject pool (Shariff and Norenzayan 2007), we hypothesized that the religious prime would increase cooperation, but only among subjects who had an experience which convinced them of the existence of God.

The results are portrayed in Fig. 2. We analyzed the data using logistic regression with robust standard errors, with PD decision as the dependent variable (0 = defect, 1 = cooperate), and prime (0 = neutral, 1 = religious) and believer (0 = does not believe in God, 1 = believes in God) as independent variables, along with a prime × believer interaction term. We also included age, gender (0 = female, 1 = male), country of residence (0 = non-U.S., 1 = U.S.) and religion (0 = non-Christian, 1 = Christian) as control variables. Consistent with our prediction, we found no significant main effect of prime ($p = 0.274$) or believer ($p = 0.545$), but a significant positive interaction between the two (coeff = 1.836, $p = 0.008$). We also found a significant main effect of gender (coeff = 0.809, $p = 0.028$), indicating that women are more likely to cooperate, but no significant effect of age ($p = 0.744$), U.S. residency ($p = 0.806$) or Christian religion ($p = 0.472$).

We demonstrated that the religious prime significantly increases cooperation in the PD, but only among subjects who had an experience which convinced them of the existence of God. These findings are of particular note given the mixed results

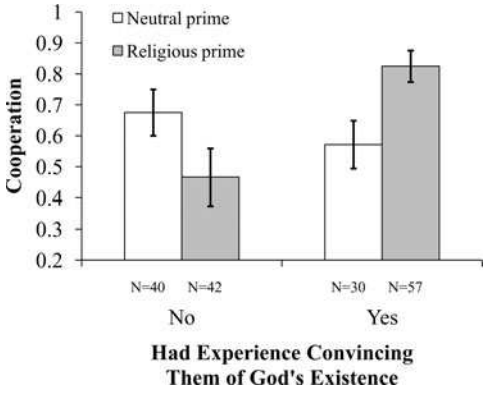


Fig. 2 Reading a religious passage significantly increases prisoner's dilemma cooperation among those who have had an experience that convinced them of the existence of God, but not among those who have not had such an experience. *Error bars* indicate standard error of the mean

of previous studies regarding the effectiveness of implicit religious primes for promoting cooperation (Benjamin et al. 2010b).[8] We demonstrate that the principle of priming can be observed with MTurk and that the effectiveness of the prime can vary systematically, depending on the characteristics of the reader.

3.4 Qualitative replication: framing

Traditional economic models assume that individuals are fully rational in making decisions—that people will always choose the option that maximizes their utility, which is wholly-defined in terms of outcomes. Therefore, decision-making should be consistent, and an individual should make the same choice when faced with equivalent decision problems. However, as the watershed experiment of Tversky and Kahneman (1981) (hereafter "TK") demonstrated, this is not the case. TK introduced the concept of "framing": that presenting two numerically equivalent situations with different language can lead to dramatic differences in stated preferences. In our current experiment, we replicated the framing effect demonstrated by TK on MTurk.[9]

In TK's canonical example, subjects read one of two hypothetical scenarios. Half of the subjects were given the following Problem 1:

> Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If Program *A* is adopted, 200 people will be saved. If Program *B* is adopted, there is $\frac{1}{3}$ probability that 600 people will be saved and $\frac{2}{3}$ probability that no people will be saved.

[8] Note that the prime had a marginally significant negative effect on subjects who had not had an experience which convinced them of the existence of God (χ^2 test, $p = 0.080$). It is conceivable that some of these subjects were antagonized by a message wrapped in religious language. The difference between the effect on believers and nonbelievers is highly significant, as indicated by the interaction term in the regression above ($p = 0.008$). The possible 'counterproductive' effect on nonbelievers may help explain the mixed results of previous studies.

[9] This is the second replication of this result on MTurk. Gabriele Paolacci also performed this experiment and reported the results on his blog, http://experimentalturk.wordpress.com/2009/11/06/asian-disease.

Which of the two programs would you favor?

The other half were given Problem 2 in which the setup (the first three sentences) was identical but the programs were framed differently:

If Program A is adopted, 400 people will die. If Program B is adopted, there is $\frac{1}{3}$ probability that nobody will die, and $\frac{2}{3}$ probability that 600 people will die.

The two scenarios are numerically identical, but the subjects responded very differently. TK found that in Problem 1, where the scenario was framed in terms of gains, subjects were risk-averse: 72% chose the certain Program A over the risky Program B. However, in Problem 2, where the scenario was framed in terms of losses, 78% of subjects preferred Program B.

Using these same prompts, we recruited 213 subjects to see whether they would reproduce this preference reversal on MTurk. We offered a participation fee of $0.40. We randomly assigned subjects to a treatment upon arrival. Consistent with TK's results, we found that the majority of subjects preferred Program A in the domain of gains ($N = 95$: 69% A, 31% B), while the opposite was true in the domain of losses ($N = 118$: 41% A, 59% B). The framing significantly affected, and in fact reversed, the pattern of preferences stated by the subjects (χ^2 test, $p < 0.001$). Thus, we successfully replicated the principle of framing on MTurk.

3.5 Field experiment: labor supply on the extensive margin

Economic theory predicts that, under most circumstances, increasing the price paid for labor will increase the supply of labor.[10] In this experiment, we exogenously manipulated the payment offered to different workers and then observed their labor supply. Because the sums involved were so small, we are confident that income effects, at least as traditionally conceived, were inconsequential in this context. We found strong evidence that subjects are more likely to work when wages are high.

When subjects "arrived" at the experiment, we explained that they would answer a series of demographic questions and then perform one paragraph-sized text transcription for a total of $0.30. They were also told that they would have the opportunity to perform another transcription after the original transcription was completed.

In addition to asking their age, gender and hours per week spent online doing tasks for money, we asked workers to identify their home countries and their primary reasons for participation on MTurk. Because economic opportunities differ by country, we might expect that motivation and behavior would also differ by country (Chandler and Kapelner 2010). Figure 3 presents a mosaic plot showing the cross-tabulation results. We can see that most subjects, regardless of nationality, claimed to be motivated primarily by money. Among those claiming some other motivation, those from India claimed to want to learn new skills, while those from the United States claimed to want to have fun.

For the actual real-effort task, we asked subjects to copy verbatim the text displayed in a scanned image into a separate text box. The text appeared as an image

[10] The exception is when the increased price changes total wealth to such an extent that changed tastes under the new scenario (i.e., income effects) might be more important than the pure substitution effect.

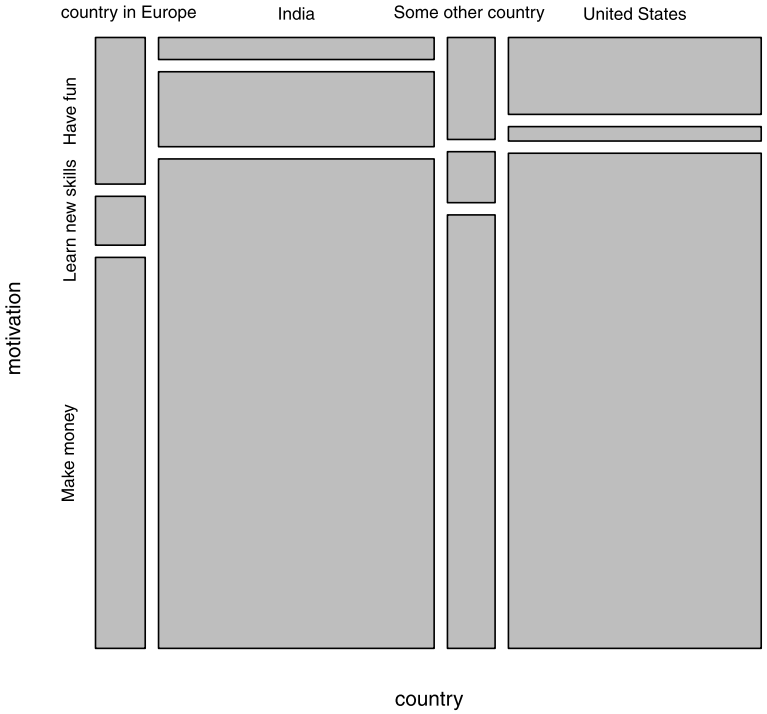


Fig. 3 Self-reported motivation for working on Amazon Mechanical Turk (*row*) cross-tabulated with self-reported country (*column*) for 302 workers/subjects

in order to prevent subjects from simply copying and pasting the text into the text box. The advantages of a text transcription task are that it (a) is tedious, (b) requires effort and attention and (c) has a clearly defined quality measure—namely, the number of errors made by subjects (if the true text is known). We have found it useful to machine-translate the text into some language that is unlikely to be familiar, yet has no characters unavailable on the standard keyboards. Translating increases the error rate by ruling out the use of automated spell-checking, and it prevents subjects from finding the true text somewhere on the web. For this experiment, our text passages were paragraph-sized chunks from Adam Smith's Theory of Moral Sentiments, machine-translated into Tagalog, a language of the Philippines.

In our experiment, after completing the survey and the first task, subjects were randomly assigned to one of four treatment groups and offered the chance to perform another transcription for p cents, where p was equal to 1, 5, 15 or 25.

As expected, workers receiving higher offers were more likely to accept. Table 1 shows that as the offered price increased, the fraction of subjects accepting the offer rose. The regression results are

$$\bar{Y}_i = \underbrace{0.0164}_{0.0024} \cdot \Delta p_i + \underbrace{0.6051}_{0.0418} \tag{3}$$

Table 1 Acceptance of paragraph transcription task by offer amount

Amount (cents)	Offers accepted	Offers rejected	Percentage accepted
1	34	37	0.48
5	57	18	0.76
15	74	5	0.94
25	71	6	0.92

with $R^2 = 0.13$ and sample size $N = 302$, with $\Delta p_i = p_i - 1$. This offsetting transformation makes the regression intercept interpretable as the predicted mean offer uptake when $p = 1$. Of course, a linear probability model only applies over a limited range, as it ultimately predicts uptake rates greater than 1. While we could use a general linear model, it makes more sense to tie the problem more closely to our theoretical model of how workers make decisions.

Presumably workers' reservation wages—the minimum amount they are willing to accept to perform some task—have some unknown distribution with cumulative density function F. Workers will choose to accept offers to do more work if the offered wages exceed their individual reservation wages. For a task taking t seconds and paying p_i cents, then $y_i = 1 \cdot \{p_i/t > \omega_i\}$, where ω_i is the reservation wage. If we assume that F is the log-normal distribution, the distribution parameters that maximize the likelihood of observing our data are $\mu = 0.113$ and $\sigma = 1.981$. Given the average completion time on the first paragraph, the median reservation wage is \$0.14/hour.

3.6 Summary

Each of these replication studies was completed on MTurk in fewer than 48 hours, with little effort required on our part. The cost was also far less than that of standard lab experiments, at an average cost of less than \$1 per subject. However, even this low per-subject cost vastly understates the comparative efficiency of online experiments. We entirely avoided both the costs associated with hiring full-time assistants and the costs of maintaining a laboratory. We also avoided the high initial costs of setting up a laboratory.

Of course, low costs would be irrelevant if the results were not informative. And yet, despite the low stakes and extreme anonymity of MTurk, the subjects' behavior was consistent with findings from the standard laboratory. The studies demonstrate the power of MTurk to quickly and cheaply give insights into human behavior using both traditional laboratory-style experiments and field experiments.

4 Internal validity

It is reassuring that our experiments achieved results consistent with those of physical laboratories, but we make an independent case for the internal validity of online experiments. Internal validity requires that subjects are appropriately assigned

to groups, that selective attrition is ruled out, and that subjects are unable either to interact with or influence one another. Each of these concerns is magnified in online settings, but fortunately there are measures that can overcome or at least mitigate these challenges.

4.1 Unique and independent observations

Because of the inherent anonymity of the Internet, would-be experimenters are rightly concerned that subjects might participate multiple times in the same experiment. Fortunately, the creators of online labor markets have their own strong financial incentives to prevent users from having multiple accounts (which would be needed for repeated participation), and all major markets use some combination of terms-of-use agreements and technical approaches to prevent multiple accounts.[11]

Our experience to date indicates they have been successful. Though vigilant to the possibility, we have detected very low numbers of people with multiple accounts (detected via use of cookies or IP address logging), though we have heard anecdotes of a few technically savvy users defeating these features.

Our view is that multiple accounts surely exist, but that they are a negligible threat in most online labor markets and are likely to remain so. As in any well-designed regulatory scheme, the steps taken by those who run the sites in online labor markets—and hence in online laboratories—raise the price of prohibited behavior. Rendering undesired behavior impossible is unrealistic, but it is possible to make such behavior highly unlikely. Furthermore, the emergence of online labor markets where subjects are non-anonymous and can easily be contacted (such as oDesk, Elance and Freelancer) provide the researcher with options to reduce the risks associated with less controlled sites like MTurk.

4.2 Appropriate assignment

To be certain that a treatment is having a causal effect on some outcome, subjects must be assigned to treatment and control groups in a way unrelated to how they will react to a treatment. Randomization is intended to meet this goal, but by chance even randomization can lead to experimental groups that differ systematically, particularly if subjects differ on characteristics strongly correlated with the outcome. One approach to this problem is to include pre-treatment variables as regressors, despite randomization. Even better, if we have complete control over assignment, as we do in the online laboratory, we can preemptively avoid the pre-treatment differences problem by using a blocking design, where we stratify on factors that correlate strongly with outcomes. Such a design creates similar groups on the basis of potentially important factors, and then applies the treatments within each group.

In online settings where subjects are recruited sequentially, we can in principle stratify subjects on any demographic characteristic we care to measure. In all of our

[11] Some sites require workers to install custom software; presumably this software can detect whether multiple copies of the software are being run on the same computer. Other sites charge membership fees or flat fees for fund transfers, both of which raise the costs of keeping multiple accounts.

experiments, we stratify according to arrival time, given the strong relationship between arrival time and demographic characteristics (driven by the global nature of online labor pools). It is important that subjects be unaware of either the stratification or the randomization; they cannot know what treatment is "next," lest it influence their participation or behavior. In principle, online experimenters could use more complicated blocking designs by adaptively allocating subjects on the basis of responses to a pre-assignment demographic survey.

4.3 Coping with attrition

Subjects might drop out of the alternative treatments in an experiment at rates that differ due to the nature of the treatments. For instance, an unpleasant video prime might prompt dropouts that a neutral video would not. Selective attrition leads to selection bias and thus poses a threat to valid inference in online experiments. The problem is especially acute online because subjects can potentially inspect a treatment before deciding whether to participate. Note that this type of balking is different from potential subjects viewing a description of the overall experiment and deciding not to participate. That is not a concern, as those people are exactly analogous to those who view an announcement for an offline, traditional experiment but don't participate. Thus, if one treatment imposes a greater burden than another, MTurk subjects are more likely to drop out selectively than their physical laboratory counterparts.

The online laboratory has at least two ways to deal with selective attrition. The first (Method 1) designs the experiment in such a way that selective attrition is highly unlikely and then shows that the data on "arrivals" is consistent with random attrition. The second (Method 2) drives down inference-relevant attrition to make it negligible. Method 1 requires the experimenter to establish the empirical fact that there is an approximate balance in collected covariates across the groups, and then to establish the untestable assumption that there is no unobserved sorting that could be driving the results (that is, different kinds of subjects are dropping out of the two groups, but by chance the total amount of attrition is the same).

It is doubtful that Method 1 could be made effectively unless the experimenter has convincing evidence from elsewhere that there are no treatment differences that could lead to disparate attrition. In our experience, such evidence is unlikely to be available: even small differences in download speed have led to noticeable and significant attrition disparities. For most applications, Method 2 is superior, though it comes at the cost of a more elaborate experimental design.

In our experience, the best way to eliminate attrition is to give subjects strong incentives to continue participating in the experiment after receiving their treatment assignments. In the physical lab, subjects will forfeit their show-up fee if they leave prematurely; this provides a strong incentive to stay. Online experiments can capitalize on something similar if they employ an initial phase—identical across treatments—that "hooks" subjects into the experiment and ensures that there is minimal attrition after the hook phase.[12] For example, all subjects might be asked to perform a rather

[12] One way of making "minimal" precise is to employ extreme sensitivity analysis and show that even all subjects selecting out had behaved contrary to the direction of the main effect, the results would still hold.

tedious but well-paid transcription task first, before being randomized to the comparatively easy tasks in the treatment and control groups. The fee for the transcription task is forfeited if the whole experiment is not completed. The experimenter then restricts the sample to subjects that persevere through the tedious first phase, which is the same for all subjects. This increases confidence that these subjects will remain for the following phase. In short, this approach has subjects invest personally in the study in a manner identical across treatment groups. We then raise the price of attrition so that any differences between treatments that might drive non-random attrition are overcome by the incentives to comply.[13]

To use this "hook" strategy ethically, it is important to let subjects know initially some of the details of the experiment. Subjects should always know approximately what they will be doing (with estimates of the time required) and the minimum payment they will receive for doing it. We have found that by providing plenty of information at the outset and by using appropriate hooking tasks, we can consistently drive attrition to zero.

4.4 Stable unit treatment value assumption

The stable unit treatment value assumption (SUTVA) requires that any individual's outcome depends only upon his or her treatment assignment, and not upon the treatment assignment or outcome of any other subject (Rubin 1974). This assumption is potentially violated if subjects can communicate with each other about their treatments, choices or experiences.

Physical laboratory experiments can avoid SUTVA problems by conducting the full experiment in one large session and prohibiting talk during the experiment. In practice, physical laboratory experiments often need to be conducted over several days to get a sufficient sample. Subjects are told not to talk to future prospective subjects. However, the extent of compliance with that request is not clear.

The SUTVA problem is both more and less challenging online than in physical laboratories. On the downside, the accumulation of subjects over time is inevitable online. The counterbalancing pluses are that the subjects are widely geographically scattered and less likely to know each other, and that the total time from first to last subject in the experiments is potentially more compressed compared to traditional field experiments, providing less time for interaction across subjects. Furthermore, SUTVA problems or their absence are likely to be known: unlike in laboratory or field experiments, the natural mode of conversations about goings-on in the market take place in publicly viewable discussion forums instead of in private encounters.

On the MTurk discussion boards, workers can and do highlight HITs that they have found particularly interesting or rewarding. Sometimes they discuss the content of the tasks. This could skew the results of an experiment. Fortunately, as an experimenter, it is easy to monitor these boards and periodically search for mentions of relevant user names or details from the experiment.

[13] Physical laboratory experiments essentially create the same pattern of costs, implying incentives not to quit. Much of the effort for participation comes from arranging a schedule and traveling to the lab.

We have been running experiments for well over a year, and occasionally search the chat rooms for mention of our user name. On one occasion, subjects did discuss our task, but a quick email from us to the original poster led to the mention being taken down. As a practical matter, we advise running experiments quickly, keeping them unremarkable and periodically checking any associated message boards for discussions of any experimental particulars.[14] Warning or threatening subjects is not recommended, as this would probably do little to deter collusion; while possibly piquing curiosity and prompting discussion.

5 External validity

As with all of experimental social science, the external validity of results is of central importance in evaluating online experiments. In this section, we discuss different aspects of external validity, including subject representativeness and quantitative versus qualitative generalizability. We also discuss the interpretation of differences between online and offline results, and present survey results comparing online and offline subjects' trust that they will be paid as described in the experimental instructions.

5.1 Representativeness

People who choose to participate in social science experiments represent a small segment of the population. The same is true of people who work online. Just as the university students who make up the subjects in most physical laboratory experiments are highly selected compared to the U.S. population, so too are subjects in online experiments, although along different demographic dimensions.

The demographics of MTurk are in flux, but surveys have found that U.S.-based workers are more likely to be younger and female, while non-U.S. workers are overwhelmingly from India and are more likely to be male (Ipeirotis 2010). However, even if subjects "look like" some population of interest in terms of observable characteristics, some degree of self-selection of participation is unavoidable. As in the physical laboratory, and in almost all empirical social science, issues related to selection and "realism" exist online, but these issues do not undermine the usefulness of such research (Falk and Heckman 2009).

5.2 Estimates of changes versus estimates of levels

Quantitative research in the social sciences generally takes one of two forms: it is either trying to estimate a level or a change. For "levels" research (for example, what is the infant mortality in the United States? Did the economy expand last quarter? How many people support candidate X?), only a representative sample can guarantee a credible answer. For example, if we disproportionately surveyed young people, we could not assess X's overall popularity.

[14] It might also be possible to "piggy-back" experiments by working with existing market participants with established, commercial reputations—an attractive option suggested to us by Dana Chandler.

For "changes" research (for example, does mercury cause autism? Do angry individuals take more risks? Do wage reductions reduce output?), the critical concern is the sign of the change's effect; the precise magnitude of the effect is often secondary. Once a phenomenon has been identified, "changes" research might make "levels" research desirable to estimate magnitudes for the specific populations of interest. These two kinds of empirical research often use similar methods and even the same data sources, but one suffers greatly when subject pools are unrepresentative, the other much less so.

Laboratory investigations are particularly helpful in "changes" research that seeks to identify phenomena or to elucidate causal mechanisms. Before we even have a well-formed theory to test, we may want to run experiments simply to collect more data on phenomena. This kind of research requires an iterative process of generating hypotheses, testing them, examining the data and then discarding hypotheses. More tests then follow and so on. Because the search space is often large, numerous cycles are needed, which gives the online laboratory an advantage due to its low costs and speedy accretion of subjects.[15]

5.3 Interpreting differences between results from online and physical laboratories

We have found good agreement between our results and those obtained through traditional means. Nonetheless, there are likely to be measurable differences between results in online and physical laboratory experiments (Eckel and Wilson 2006). How should one interpret cross-domain differences, assuming they appear to be systematic and reproducible?

First, systematic differences would create puzzles, which can lead to progress. In this sense, the online laboratory can complement conventional methods—a point emphasized by Harrison and List (2004) in the context of comparing laboratory and field experiments. To give an example, suppose that future experiments find that subjects donate less in an online dictator game than in person. This would tell us something interesting and add to our knowledge. It would not mean that the Internet result is "wrong," assuming that the game is set up properly. The situation is analogous to finding cross-cultural differences in game play (Bohnet et al. 2008; Gneezy et al. 2009; Herrmann and Thöni 2009). Such results increase our knowledge of the social world; they are not cautionary tales about the desirability of restricting experimental subjects to undergraduates at U.S. research universities.

When the Internet was small and few people spent much time online, perhaps it made sense to treat cross-medium differences as an argument against utilizing evidence from online domains. Today, however, Internet-mediated social interactions are no longer a strange experience, one familiar to only a tiny fraction of the population.[16]

[15] It also increases the danger of "hypothesis mining"—trying out many hypotheses and reporting those that work, quite likely only by chance.

[16] A recent *New York Times* article, "If Your Kids Are Awake, They're Probably Online," reported that American kids spend an average of seven to eight hours per day online.

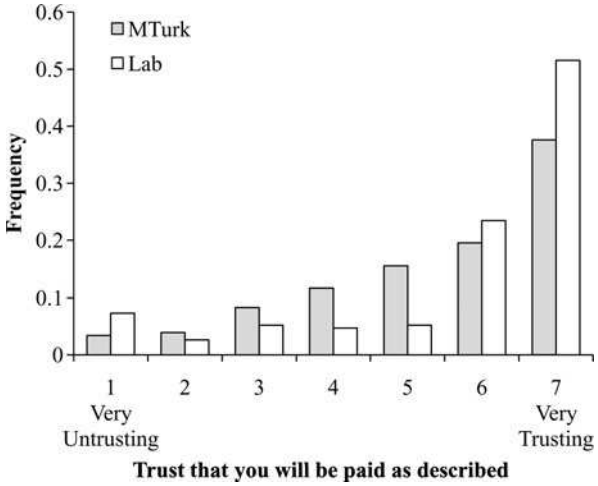


Fig. 4 Degree of stated trust in experimenter instructions regarding payoffs among MTurk workers and subjects recruited from the Harvard Decision Sciences Laboratory

5.4 Comparison of beliefs of online and offline subjects

In economic experiments, participants must believe that payoffs will be determined as the experimenter describes. We wished to assess how MTurk workers compare to participants in traditional lab experiments in their level of experimenter trust. Thus, we conducted a survey receiving responses from 205 workers on MTurk and 192 members of the Harvard Decision Sciences Laboratory subjects pool (a lab that prohibits deception).[17] Participants indicated the extent to which they trusted that they would be paid as described in the instructions using a 7-point Likert scale, with 7 being the maximum trust level.

Responses are shown in Fig. 4. Although there is lower trust among MTurk workers, the mean level of trust is comparable and highest trust response was modal for both groups. The mean levels of trust were 5.41 for MTurk workers and 5.74 for lab subjects. The difference was 0.19 standard deviations, and represented only a modest difference in trust levels, though one that was strongly statistically significant using a rank-rum test ($p = 0.004$). These survey results indicate that trust of experimenters is fairly high on average. The physical lab had more "extreme" subjects—very trusting and very untrusting—and fewer in the middle.

6 Experimental designs

Certain kinds of research designs work better than others online. One design that works well is the experimenter-as-employer natural field experiment. For such experiments, the interaction is completely unremarkable from the perspective of online subjects, who have no idea they are involved in an experiment. Of course, for this kind of experiment, the researcher's institutional review board (IRB) must waive the

[17] MTurk subjects were paid \$0.10 for responding; lab pool subjects received no payment.

need to obtain informed consent from subjects prior to participation. Institutions vary in their requirements, but we are hopeful that boards will approve experimenter-as-employer studies where the task and payments fall within the range of what normally occurs on the labor market, as they often do for traditional, offline field experiments.

Certain kinds of surveys also work well online, as we will discuss below. Laboratory-type games are certainly possible, but they have a number of limitations, as discussed above, some of which will likely be overcome by better software, while others will remain intractable due to the inherent nature of the Internet.

6.1 Experimenter-as-employer

In many offline field experiments, the researcher participates in the market as an employer, either by creating a new "firm" or by piggy-backing on an existing firm and directing its policies. Worker-subjects perform a real-effort task, such as stuffing envelopes, picking fruit, planting trees, etc. The manipulation is usually of such elements as the payment scheme, team composition or framing. The online environment makes it easy to conduct experiments involving such manipulations, so long as the task requires no physical presence. Fortunately, there are plenty of tasks that meet this requirement, such as data entry, content creation and image labeling.

Subjects are simply recruited from the existing market to perform a task, one equivalent to many tasks required by real employers. They are then randomly assigned to groups. The objective might be to test the role of alternative incentive schemes or to determine how payment affects quality of performance. Depending upon institutional review board requirements, the subjects might not need to be notified that the task is an experiment.

For the actual online task, it is obviously impossible to have workers perform something physical, yet certain real-effort tasks are well-suited to online completion. We have used data entry tasks (namely transcribing text from a document) and image labeling tasks. Both are common tasks on MTurk, and are easily understood by workers. The advantage of text transcription is that it has an unambiguous error metric if the true text is known. The advantage of labeling is that if subjects can decide how many labels to produce, creating the variation in the output measure needed for many experiments.

The paired survey represents a variant of the experimenter-as-employer paradigm. It is used when one wants to know how some feature of the way a question is posed affects responses. While this kind of experiment is very simple, it can yield powerful conclusions. For example, some of Tversky and Kahneman's classic work on framing, which we replicated, used the paired-survey method to show that people viewed objectively identical scenarios quite differently depending upon whether a two-component outcome was framed to focus on the gain or loss component. This simple survey design yielded an insight that has subsequently been shown to be widespread and important. MTurk is ideal for this kind of research.

6.2 Laboratory-type games online

Some of the most successful examples of experimental social science use simple interactive games, such as the trust game, the prisoner's dilemma and market games.

Subject interactions are easy to arrange in physical laboratory experiments, because all the subjects play at once. In online labor markets, by contrast, subjects often arrive at the experiment over the course of several hours or days, making subject interactions difficult. There are several solutions to the difficulty presently available and many more under development.

When subjects are asynchronous, the widely used strategy method (Selten 1967)—players report what they would do in various hypothetical situations—can identify outcomes in interactive situations. This was the method we employed when performing our own trust-game and ultimatum-game experiments (not yet published). There is some evidence that subjects play "hot" games (those that do not use the strategy method) differently (Brandts and Charness 2000). Ideally, new software developments will allow for hot interactive games.

If the reliability of the strategy method is accepted, implementation of online experiments is simple. The experimenter needs only to simulate play once all responses have been received. This method also has the advantage of giving more data points. For example, in contrast to a "hot" ultimatum game where we can only observe "accept" or "reject" responses from the second player, in the "cold" strategy-method game, we can see the subjects' responses to several offers because they must answer the question "Would you accept X?" for several X values.

The online laboratory allows for "hot" play if sufficiently large numbers of respondents can be matched up as they arrive. Experiments have shown that it is possible to get MTurk workers to wait around for another player.[18] This kind of approach requires that the experimenter establish some rules for payment if a match cannot be found in a suitable amount of time. Another approach is to pay workers to come back to a website at a certain time to play the game. The great advantage of this method is that it closely replicates the current laboratory experience. This method requires web-based interfaces for games. Work at MIT to develop "Seaweed," a web-based experimental platform, represents a strong step in this direction (Chilton et al. 2009). Building such platforms should be a top priority for the experimental social science community.

7 Ethics and community

The online laboratory raises new ethical issues. However, it also creates opportunities for setting higher standards of evidence and for fostering greater collaboration among social scientists, both in terms of sharing materials and in the adversarial "collaboration" of replication studies, which are now far easier to perform.

7.1 Ethical implications of moving towards a bench science

Online experiments can now propel certain subfields in the social sciences substantially toward "bench science." It is now remarkably easy to design, launch and analyze

[18] See this blog post report of the experiment by Lydia Chilton: http://groups.csail.mit.edu/uid/deneme/?p=483.

the results of an experiment. Conducting multiple experiments per week for relatively small amounts of money is now feasible. While this is an exciting and welcome development, in such an environment, a researcher could turn out a stream of spuriously significant results by burying all that are negative. Even honest researchers can convince themselves of the flaws in their "pilots" and of the legitimacy of the subsequent experiments that happened to yield good results.[19]

A significant advantage of online experiments is that they can be run with little assistance from others. This creates the offsetting disadvantage of reducing critiques by others of procedures and results. Since there are no lab technicians, no subjects who can be contacted and no logs on university-run servers, the temptation to cheat may be high. While few researchers would knowingly falsify results, certain professional norms could raise the cost of bad behavior, with the effect of both fostering honesty and dampening skepticism.

The first norm should be to require machine-readable sharing of experimental materials, as well as of detailed instructions on set-up and process. Perhaps some institution, such as a professional organization or the National Science Foundation, could set up a library or clearinghouse for such materials. While most results would probably not be checked by new experiments, requiring all experimenters to make replication very easy would make all results "contestable." This should help make cheating an unprofitable and unpopular strategy. Another advantage of such a norm is that it would reduce costly duplication of programming effort and design. As a current example, the open-source survey software Limesurvey allows researchers to export their survey designs as stand-alone files. Researchers can simply download other people's experimental materials and then deploy their own working versions of surveys/experiments.

Within economics, a consensus is developing to make all data and code publicly available. To adhere to and support this norm is easy in online contexts, but online experimenters should go a step further. Datasets should be publicly available in the rawest form possible (that is, in the format in which the experimental software collected the data), as should the associated code that turned the raw data into the data set.[20] The Internet makes such sharing a low-cost chore, since the data are invariably generated in machine-readable form.

7.2 Deception

Experimental economics has a well-established ethic against deceiving subjects, an ethic that provides significant positive externalities. Many experiments rely critically on subjects accepting instructions from experimenters at face value. Moreover, deception in money-staked economics experiments could approach and even constitute

[19] It is thus important that any final paper describe the alternative formulations that were tried. Statistical tests should take alternatives tried into account when computing significance.

[20] Often it is necessary to clean this data in different ways, such as by dropping bad inputs, or adjusting them to reflect the subject's obvious intent (e.g., if a subject is asked to report in cents and reports .50, it might reasonable to infer they meant 50 cents, not a half-cent). By making all trimming, dropping, reshaping, etc., programmatic, it is easier for other researchers to identify why a replication failed, or what seemingly innocuous steps taken by the original researcher drove the results.

fraud. The arguments for maintaining this "no-deception" policy in online experiments are even stronger.

Workers in online labor markets represent a common resource shared by researchers around the globe. Once experiments reach a significant scale, practicing deception in these markets could pollute this shared resource by creating distrust. In the online world, reputations will be hard to build and maintain, since the experimenter's user name will usually be the only thing the subject knows about the experimenter. Of course, economists will share the experimental space with psychologists, sociologists and other social scientists who may not share the "no deception" norm. Probably little can be done to police other disciplines, but economists can take steps to highlight to their subjects that they adhere to the no-deception rule and to present arguments to others that deception is a threat to the usefulness of these markets. If online experiments become truly widespread, we would expect some sites to prohibit deception, in part because their non-experimenting employers would also be hurt by distrust. Additional forms of certification or enforcement processes are also likely to arise.

7.3 Software development

The most helpful development in the short term would be better (and better-documented) tools. There are a number of software tools under active development that should make online experimentation far easier. The first goal should probably be to port some variant of zTree to run on the Internet. The MIT initiative "Seaweed" is a nascent attempt at this goal, but it needs much more development.

Obviously some tools, such as those for complex games, will have to be custom-built for individual experiments. Advances in software are required for games calling for simultaneous participation by two or more subjects. Such software is already under development in multiple locales. However, the research community should, as much as possible, leverage existing open-source tools. For example, many experiments simply require some kind of survey tool. The previously mentioned open-source project, "Limesurvey," already offers an excellent interface, sophisticated tools and, perhaps most importantly, a team of experienced and dedicated developers as well as a large, non-academic user base.

8 Conclusion

We argue in this paper that experiments conducted in online labor markets can be just as valid as other kinds of experiments, while greatly reducing cost, time and inconvenience. Our replications of well-known experiments relied on MTurk, as MTurk is currently the most appropriate online labor market for experimentation. However, as other online labor markets mature and add their own application programming interfaces, it should be easier to conduct experiments in additional domains. It might even be possible to recruit a panel of subjects to participate in a series of experiments. While MTurk workers remain anonymous, many other markets have the advantage of making it easier to learn about subjects/workers.

We have shown that it is possible to replicate, quickly and inexpensively, findings from traditional, physical laboratory experiments in the online laboratory. We have also argued that experiments conducted in the context of online labor markets have internal validity and can have external validity. Lastly, we have proposed a number of new and desirable norms and practices that could enhance the usefulness and credibility of online experimentation. The future appears bright for the online laboratory, and we predict that—as the NetLab workshop quotation in our introduction suggests—the social sciences are primed for major scientific advances.

References

Andreoni, J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *The Economic Journal*, 464–477.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390.

Bainbridge, W. S. (2007). The scientific research potential of virtual worlds. *Science*, *317*(5837), 472.

Benjamin, D. J., Choi, J. J., & Strickland, A. (2010a). Social identity and preferences. *American Economic Review* (forthcoming).

Benjamin, D. J., Choi, J. J., Strickland, A., & Fisher, G. (2010b) *Religious identity and economic behavior*. Cornell University, Mimeo.

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal aversion: evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, *98*(1), 294–310.

Brandts, J., & Charness, G. (2000). Hot vs. cold: sequential responses and preference stability in experimental games. *Experimental Economics*, *2*(3), 227–238.

Camerer, C. (2003). *Behavioral game theory: experiments in strategic interaction*. Princeton: Princeton University Press.

Chandler, D., & Kapelner, A. (2010). *Breaking monotony with meaning: motivation in crowdsourcing markets*. University of Chicago, Mimeo.

Chen, D., & Horton, J. (2010). *The wages of pay cuts: evidence from a field experiment*. Harvard University, Mimeo.

Chilton, L. B., Sims, C. T., Goldman, M., Little, G., & Miller, R. C. (2009). Seaweed: a web application for designing economic games. In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 34–35). New York: ACM Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Boston: Houghton Mifflin.

Eckel, C. C., & Wilson, R. K. (2006). Internet cautions: experimental games with Internet partners. *Experimental Economics*, *9*(1), 53–66.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535.

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

Fehr, E., Schmidt, K. M., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994.

Fischbacher, U. (2007). z- tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Frei, B. (2009). *Paid crowdsourcing: current state & progress toward mainstream business use*. Produced by Smartsheet.com.

Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: evidence from a matrilineal and a patriarchal society. *Econometrica*, *77*(5), 1637–1664.

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

Herrmann, B., & Thöni, C. (2009). Measuring conditional cooperation: a replication study in Russia. *Experimental Economics*, *12*(1), 87–92.

Horton, J. (2010). Online labor markets. In *Workshop on Internet and network economics* (pp. 515–522).

Horton, J. (2011). The condition of the Turking class: are online employers fair and honest? *Economic Letters* (forthcoming).

Horton, J. & Chilton, L. (2010). The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on electronic commerce*.
Ipeirotis, P. (2010). *Demographics of Mechanical Turk*. New York University Working Paper.
Kagel, J. H., Roth, A. E., & Hey, J. D. (1995). *The handbook of experimental economics*. Princeton: Princeton University Press.
Kittur, A., Chi, E. H., & Suh, B. (2008). *Crowdsourcing user studies with Mechanical Turk*.
Kocher, M. G., & Sutter, M. (2005). The decision maker matters: individual versus group behaviour in experimental beauty-contest games*. *The Economic Journal*, *115*(500), 200–223.
Levitt, S. D., & List, J. A. (2009). Field experiments in economics: the past, the present, and the future. *European Economic Review*, *53*(1), 1–18.
Little, G., Chilton, L. B., Goldman, M., & Miller, R. C. (2009). TurKit: tools for iterative tasks on Mechanical Turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. New York: ACM Press.
Lucking-Reiley, D. (2000). Auctions on the Internet: what's being auctioned, and how? *The Journal of Industrial Economics*, *48*(3), 227–252.
Mason, W., & Watts, D. J. (2009). Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 77–85). New York: ACM Press.
Mason, W., Watts, D. J., & Suri, S. (2010). *Conducting behavioral research on Amazon's Mechanical Turk*. SSRN eLibrary.
Pallais, A. (2010). *Inefficient hiring in entry-level labor markets*.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*.
Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45–48.
Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: a controlled experiment. *Experimental Economics*, *9*(2), 79–101.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.
Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschrankt rationalen Verhaltens im Rahmen eines Oligopolexperiments. *Beitrage zur experimentellen Wirtschaftsforschung*, *1*, 136–168.
Shariff, A. F., & Norenzayan, A. (2007). God is watching you. *Psychological Science*, *18*(9), 803–809.
Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 614–622). New York: ACM Press.
Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. *University of Illinois at Urbana-Champaign, Mimeo*, *51*, 61820.
Suri, S., & Watts, D. (2011). A study of cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE* (forthcoming).
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453.
von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003). CAPTCHA: using hard AI problems for security. In *Lecture notes in computer science* (pp. 294–311). Berlin: Springer.

Economics Letters 111 (2011) 10–12

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

The condition of the Turking class: Are online employers fair and honest?

John J. Horton *

Harvard University, 383 Pforzheimer Mail Center, 56 Linnaean Street, Cambridge, MA 02138, United States

ARTICLE INFO

Article history:
Received 19 January 2010
Received in revised form 11 November 2010
Accepted 1 December 2010
Available online 10 December 2010

JEL classification:
J4
J8
D6

Keywords:
Crowdsourcing
Experimentation
Amazon mechanical turk
Online labor markets

ABSTRACT

Critics of online labor markets claim that employer abuses are endemic in these markets. Surveying a sample of workers, I find that, on average, workers perceive online employers to be slightly fairer and more honest than offline employers.

1. Introduction

Work conducted over the Internet by workers participating in online labor markets has begun to attract mainstream attention, much of which has been negative. Most of the criticism targets low-skilled, low-paying piece-work sites like Amazon's Mechanical Turk (MTurk), though other remote work sites like oDesk and Elance are also drawing scrutiny [2]. Some critics worry that much of the work is of dubious social value, with many employers using workers to generate spam and write bogus product reviews. Other critics argue that buyers are actively exploiting workers and circumventing labor laws. In a recent *Newsweek* article entitled "Work the New Digital Sweatshops," Harvard Law School Professor and Berkman Center for Internet & Society co-founder Jonathan Zittrain characterized online labor markets as digital sweatshops. At a recent conference on digital labor at the New School, expropriation and exploitation were common themes. Aside from low pay, critics argue that workers do not know the (potentially unethical) purpose of their work and have no ability to organize or appeal the decisions of capricious employers [3,5].

Despite these real and perceived downsides, online labor markets have a tremendous and potentially transformative upside, which is that the markets give people in poor countries access to buyers in rich countries. If this form of increased virtual labor mobility has effects similar to those of increased real labor mobility, then the emergence of online labor markets should be lauded and supported; the welfare gains from liberalizing restrictions on labor mobility would be truly enormous. Clemens et al. [1] consider the effect relocation to the US would have on the *real* wages of workers from different countries: for the median country (Bolivia), wages would increase by a factor of 2.7, and for the highest country (Nigeria), wages would increase by a factor of 8.4. Even with current strict limits on migration, the World Bank estimates that in 2008 remittances to developing countries totaled over $305 billion, exceeding both private capital flows and official development aid.

The comparative advantages of the world's poor are that they (either individually or collectively through political institutions) are willing to accept environmental degradation, dangerous working conditions and very low pay. In light of this unpleasant truth, the relative virtues of digital work are obvious: it poses no physical danger to workers, has virtually no environmental impact and does not require robust host country institutions or local entrepreneurial talent. Workers can set their own hours and are not exposed to the elements, dangerous working conditions, the vagaries of agriculture or tyrannical bosses. Unlike labor market access gained through physical migration, workers do not have to live apart from their families or dissipate their earnings by paying developed-country prices for shelter, food and clothing.

The perceived costs and the potential benefits of online labor markets are fairly self-evident; good public policy will require some attempt to quantify the trade-offs under different policy scenarios. Consider a proposal to require Amazon to verify that each new task is not being used

* Tel.: +1 617 595 2437.
E-mail address: john.joseph.horton@gmail.com.

0165-1765/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2010.12.007

to generate spam. While this may have the intended effect of reducing spam, compliance costs might exceed the per-transaction profits, thus pricing all work out of the market — or it might not, leading to an overall gain in welfare. Fortunately, we do not need to limit ourselves to speculation and conjecture: it is easy to test hypotheses about online phenomena by running experiments.

2. The experiment

In MTurk, the decision whether to pay a worker is left to the employer's discretion; nothing prevents a buyer from expropriating a worker's product. One might therefore presume that employers in MTurk would regularly cheat their workers. If this is true, then MTurk workers should have dim views of the honesty and fairness of online employers compared to offline employers who operate under a formal framework of sanctions for unethical behavior. To test this proposition, I conducted a simple experiment in which subjects recruited from MTurk were asked to answer one of the following questions:

- What percentage (between 0 and 100) of employers in your home country would you estimate treat workers honestly and fairly?
- What percentage (between 0 and 100) of Mechanical Turk Requesters would you estimate treat workers honestly and fairly?

Subjects were randomly assigned to a question and only saw that particular question. In other words, subjects asked about home country employers were not asked about online employers, and vice versa. I launched the experiment on December 23, 2009, and left it open for 7 days. Exactly 200 subjects completed the HIT, but only 192 responses were usable.[1] I paid 12 cents per response, which gave an hourly average wage of $5.68, and my total expenditure was $26.40. Consistent with Amazon's guidance on what constitutes a "good" feedback score, the HIT was limited to MTurk workers with a 95% approval rating. When asked to name their home countries, 111 subjects reported the US, 58 reported India and 23 reported some other country.

2.1. Results

A histogram of worker responses by question type, with a bin width of 5, is shown in Fig. 1. In each panel, the mean response is indicated with a vertical line, and 90% and 95% confidence intervals are shown with red shaded bands around the mean. We can see that (1) means are quite similar and (2) responses in the online case (bottom panel) are slightly more polarized, with a greater number of subjects reporting very positive views of online employers.

We can compare mean responses in the two groups using a linear regression of the reported percentage of honest and fair employers, $perc_i$, on an indicator for whether the subject was asked about online employers, $MTurk_i = 1$ or offline, host country employers, $MTurk_i = 0$. The fitted regression line, with robust standard errors, is:

$$perc_i = \underset{[3.58]}{5.21} \cdot MTurk_i + \underset{[2.36]}{64.38}$$

with $R^2 = 0.011$ and $N = 192$. We can see that the mean percentage for offline employers was a little more than 64% and for online employers slightly more than 69%. The difference is not statistically significant.

Confirming the graphical evidence that more workers have very positive views of online employers compared to offline employers, a regression of an indicator for whether the subject perceived that more than 80% of the employers were honest and fair yields a positive and

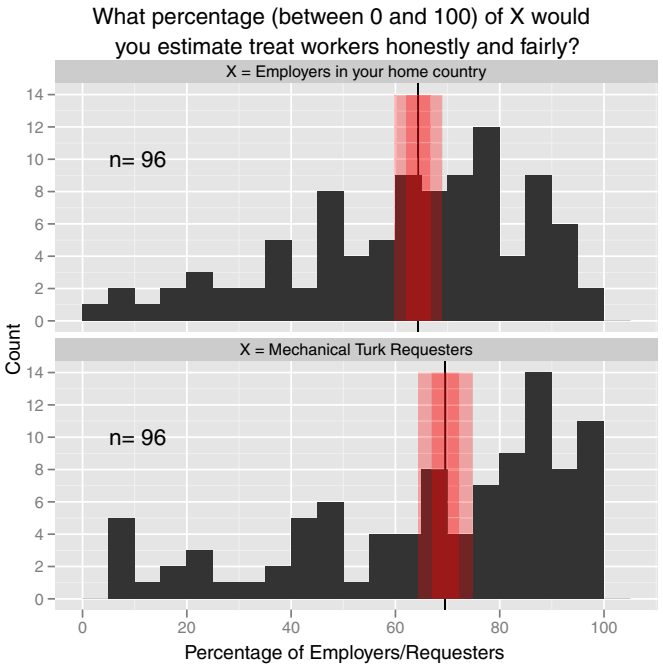


Fig. 1. Worker attitudes about online and offline employers.

highly significant coefficient on the group assignment indicator, $MTurk_i$:

$$1\{perc_i > 80\} = \underset{[0.07]}{0.22} \cdot MTurk_i + \underset{[0.04]}{0.22}$$

with $R^2 = 0.054$ and $N = 192$.

Perhaps surprising, given the cross-national differences in institutions, Indian and non-Indians (composed mostly of US-based workers) had very similar responses:

$$perc_i = -\underset{[5.57]}{3.00} \cdot MTurk_i + \underset{[4.32]}{4.71} \cdot India_i + \underset{[7.95]}{1.81} \cdot India_i \times MTurk_i + \underset{[2.71]}{65.25}$$

with $R^2 = 0.013$ and $N = 192$.

3. Discussion

These findings come with several caveats. Experimenter effects could matter — subjects might be encouraged to exaggerate how honest and fair they find AMT employers, because the question was asked by an MTurk employer. An unavoidable limitation is that subjects are not a random sample of MTurk workers — the experiment cannot reach workers who may have stopped using MTurk because of bad experiences. Additionally, the 95% cutoff might preclude the participation of a large number of disgruntled workers, though this seems unlikely; in past experiments, I have found very little difference in uptake under different cutoffs, suggesting that most workers have high scores.

The critique of online labor markets goes beyond the perceived fairness of employers. Furthermore, worker perceptions of fairness are not measures of *actual* fairness. Nevertheless, this experiment offers evidence that MTurk workers view their chances of being treated fairly online as being as good or better than offline. Contrary to our prior expectations, rampant exploitation is a mis-characterization.

[1] The author's website contains the raw data from MTurk, the code that cleans the data and all code used for the analysis and plots. The main cause was workers offering a range rather than a point estimate.

Future research should investigate other claims related to online markets: How prevalent are dubious tasks? Do workers get repetitive stress injuries, as is often suggested? Do workers feel they are gaining skills? Answers to these questions could help clarify the trade-offs inherent in different policy proposals. Online work is currently a small phenomenon compared to the global trade in services, but it will become far larger and will eventually attract more policy-oriented attention. Given the welfare consequences of online work, it would be a tragedy if supposition and conjecture about easily and cheaply answerable empirical questions determined our policy towards digital markets.

Acknowledgments

John Horton thanks the Berkman Center for Internet and Society and the NSF-IGERT Multidisciplinary Program in Inequality & Social Policy for generous financial support. Thanks to Aaron Shaw, Roberta Yerkes Horton, Justin Keenan, Richard Zeckhauser, Carolyn Yerkes and Fara Dina Aquino for helpful comments and suggestions. All plots were made using the open source R package `ggplot2`, developed by Hadley Wickham [4].

References

[1] Clemens, M., Montenegro, C., Pritchett, L., Paraguay, D., Floor, T., 2008]. The place premium: wage differences for identical workers across the US border. World Bank Policy Research Working Paper No. 4671.
[2] Felstiner, A., 2010. Working the crowd: Emplyment and labor law in the crowdsourcing industry. Working Paper, Boalt Hall Law School, UC Berkeley.
[3] Irani, L., 2009]. Agency and exploitation in mechanical turk. Internet as Playground and Factory Conference. New School, New York, NY. Nov 13.
[4] Wickham, H., 2008]. ggplot2: an implementation of the grammar of graphics. R package version 0.7, URL: http://www.CRAN.R-project.org/package=ggplot22008.
[5] Zittrain, J., 2008]. Ubiquitous human computing. Philosophical Transactions of the Royal Society 3813–3821.

Labor Allocation in Paid Crowdsourcing: Experimental Evidence on Positioning, Nudges and Prices*

Dana Chandler and **John Horton**
MIT
oDesk Corporation

Abstract

This paper reports the results of a natural field experiment where workers from a paid crowdsourcing environment self-select into tasks and are presumed to have limited attention. In our experiment, workers labeled any of six pictures from a 2 x 3 grid of thumbnail images. In the absence of any incentives, workers exhibit a strong default bias and tend to select images from the top-left ("focal") position; the bottom-right ("non-focal") position, was the least preferred. We attempted to overcome this bias and increase the rate at which workers selected the least preferred task, by using a combination of monetary and non-monetary incentives. We also varied the saliency of these incentives by placing them in either the focal or non-focal position. Although both incentive types caused workers to re-allocate their labor, monetary incentives were more effective. Most interestingly, both incentive types worked better when they were placed in the focal position and made more salient. In fact, salient non-monetary incentives worked about as well as non-salient monetary ones. Our evidence suggests that user interface and cognitive biases play an important role in online labor markets and that salience can be used by employers as a kind of "incentive multiplier".

Introduction

In this paper, we conduct a *natural field experiment* (Harrison and List 2004) in a paid crowdsourcing environment to study how worker decision making is affected by the type and saliency of incentives. In particular, we compare the effects of direct monetary incentives (which workers presumably directly value) to non-monetary incentives that merely indicate employer preferences without directly affecting workers' material payoffs. Further, we investigate how these two types of incentives interact both with each other and with worker default behaviors (related to the pattern by which they scan a document or interface) that can give certain kinds of incentives more or less salience.

In Section we discuss the conceptual framework needed to understand the experiments and the theory that motivates our designs. Section explains the experimental design, Section reports the results and Section concludes.

Conceptual Framework

The power of positional effects has been extensively studied by human factors researchers; one well-established regularity is that people who read "left-to-right" in their native language tend to scan information in a corresponding left-to-right and top-to-bottom manner. This tendency has been well-documented in the context of navigating web pages and selecting items from computer interfaces (Goldberg et al. 2002; Hornof 2004). Positional effects also play an important role in how workers search for tasks in Amazon's Mechanical Turk (MTurk) marketplace: users overwhelmingly tend to select tasks that are positioned first(Chilton et al. 2010). These positional effects suggest that people evaluate their options with time-saving heuristics of "default" behaviors.

Heuristics, attention and salience

Because it is not free to gather or process information, rational agents frequently employ shortcuts and thus often *rationally* decide to overlook certain aspects that may be relevant to a decision.[1] In the real world where people have limited attention, it matters how choices are presented.

Several empirical papers have documented that people respond more to *salient* prices. In toll booths where tolls are collected electronically, the price of tolls tend to rise more rapidly (even after controlling for confounding factors) suggesting that toll operators raise prices more since drivers are less aware of price increases when they do not physically take money out of their pockets to pay tolls (Finkelstein 2009).[2] In the public policy domain, Thaler and Sunstein argue in their book *Nudge: Improving Decisions About Health, Wealth, and Happiness* that, "setting default options, and other similar seemingly trivial menu-changing strategies, can have huge effects on outcomes, from increasing savings to improving health care to providing organs for life-saving transplant operations." (Thaler and Sunstein 2008).

[1] Herbert Simon, extended the concept of "homo economicus" with a model of "bounded rationality" in his seminal 1955 article (Simon 1955).

[2] DellaVigna also provides an overview of other interesting economic field experiments that measure the impact of limited attention, salience, and other cognitive biases (DellaVigna 2009)

Figure 1: The 2 x 3 grid of images from our task interface

Notes: This figure shows the 2 x 3 layout used in our experiment. Depending on treatment, different levels of bonuses and progress bars were shown below each image. Above, we show Treatment 6 where the bottom-right position has a 5-cent bonus and a 10% progress bar as compared with a 1-cent bonus and 90% progress bar in every other position.

Experimental approach

We examine the consequences of default behavior and investigate whether workers persist in picking whichever image arbitrarily appears in the top left, even when the images are made to have different payoffs.

Our first series of experiments uses *progress bars* to indicate that images have different levels of completion and examines whether workers will choose the task that is most in need of completion. Although workers are provided with an indication of how close to completion the work is, they are *not* rewarded for selecting any particular image.

Our second series of experiments parallels the first, but instead of progress bars, the experiment uses *bonuses* between 1 and 5 cents in order to observe whether workers will choose the task that gives them the highest explicit monetary incentive.

Given the importance of default behavior and limited attention, we also test whether the *position*—i.e., salience—of each incentive type also has an effect. Several of our treatments are identical except that they are placed in either a "focal" position (i.e., in the top-left) or a "non-focal" position (i.e., the bottom-right); by holding the incentive types constant and varying position, we can identify the effect of salience. A full description of both sets of treatments can be found in section .

To summarize, our treatments were divided along the following dimensions:

- **type of incentive**: monetary, progress bars or both
- **size of monetary incentive**: large bonus (5-cents) or small bonus (1-cent)
- **saliency/position of incentive**: focal or non-focal

Experimental Design

Subjects were recruited from Amazon Mechanical Turk (MTurk) and were offered a payment of 25 cents to label a single image. We recruited a total of 837 subjects. The subjects were stratified over the seven treatment groups.

Our job posting was designed to appear just like any other image-labeling task that workers ordinarily perform in this labor market. As a result, the behavior we observe is likely to be representative of ordinary behavior in this marketplace. This design is known as a "natural field experiment" (Harrison and List 2004) that uses the "experimenter-as-employer" paradigm(Gneezy and List 2006; Horton, Rand, and Zeckhauser forthcoming) [3] (Mason and Watts 2009) provides an overview of financial incentives on MTurk and (Shaw, Horton, and Chen 2010) provides an overview of a variety of non-monetary incentives.

Description of task and task interface

In our experiment, workers/subjects were asked to provide labels for photographic images. This is a common task in the MTurk market, as the challenges of computer vision make image labeling a prototypical human computation task.(Huang et al. 2010; Von Ahn and Dabbish 2004)

Before beginning, workers were given treatment-specific instructions on how to complete their task. Afterwards, they were presented with a screen showing a 2 x 3 array of images where they chose an image to label. Figure 1 shows the array and one particular configuration of our experimental

[3] Other recent examples that use this design include (Chandler and Kapelner 2010; Horton and Chilton 2010; Shaw, Horton, and Chen 2010; Mason and Watts 2009; Pallais 2010).

parameters (i.e, progress bar, bonus and position). Images were randomly assigned to the six positions and randomized for each trial.[4]

Our primary outcome measure was the position a worker chose to label from a 2-row, 3-column array of thumbnail images.

Types of incentives

Progress bars Treatments 1, 2 and 3 examine whether an arrangement of progress bars affected task selection. Subjects were told that:

> *"[progress bars] indicate how near we are to our labeling goals for each image... 100% indicates that we have reached our goal and don't need any more labels. 0% indicates that we have no labels for that image."*

Table 1: Treatments involving progress bars

Treatment Number	Name	Level of progress bars	Position of 10% bar
0 ($n = 118$)	Balanced	All at 10%	–
1 ($n = 117$)	Imbalanced/ Non-salient	5 at 90%, 1 at 10%	bottom-right
2 ($n = 115$)	Imbalanced/ Salient	5 at 90%, 1 at 10%	top-left

Table 1 lists all of our *progress bar* treatments.

Monetary incentive (bonus) treatments Our remaining treatments offered workers monetary bonuses in exchange for labeling images at certain positions. Each position was given a particular bonus level; a worker selecting an image received the bonus associated with that position (if any). Across treatments, we varied the size of the bonuses, the placement of the bonuses and whether bonuses were paired with progress bars.

Table 2: Treatments involving bonuses

Treatment Number	Name	Level of bonuses	Position of largest bonus
3 ($n = 125$)	Large Bonus/ Non-salient	5 at 1c, 1 at 5c	bottom-right
4 ($n = 125$)	Large Bonus/ Salient	5 at 1c, 1 at 5c	top-left
5 ($n = 109$)	Small Bonus/ Non-salient	5 at 0c, 1 at 1c	bottom-right
6 ($n = 128$)	Large Bonus/ Salient	5 at 1c, 1 at 5c	bottom-right

[4]This allows us to disentangle whether a particular *position* was more popular and whether a particular *image* was more popular.

Design and predictions from bonus treatments Across our bonus treatments, we paid two levels of bonuses: a five cent ("large bonus") and a 1 cent ("small bonus"). In Treatment 3, a 5-cent large bonus was placed on the lower right hand corner—the non-focal position. Every other position offered a 1-cent small bonus. In Treatment 4, a 5-cent bonus was placed on the focal position while every other position offered a 1-cent bonus. (Note the parallelism with Treatments 1 and 2.)

We next examined whether the size of the bonus matters. If workers do indeed have lexicographic preferences, they should select the image that offers the highest bonus, even if it is only 1-cent. Treatment 5 tested this by offering a 1-cent bonus in the non-focal position and no bonus (0 cents) at every other position.

In Treatments 3, 4 and 5, the levels of the progress bars were set to be neutral; specifically, all progress bars were set at 10% and the only dimensions of the experiment that changed were the size and position of the bonuses. In contrast, Treatment 6 used imbalanced progress bars, implicitly suggesting the non-focal image. Treatment 6 measured the double effect of using both progress bars and bonuses to draw people towards the bottom right. In essence, Treatment 6 combines Treatments 1 and 4 by employing both an imbalanced configuration of progress bars and a large (5-cent) bonus in the non-focal position.

Table 2 summarizes all of the *bonus* treatments.

Experimental results

In the subsections below, we present results relating to a number of hypotheses. The raw data for each of the seven experimental groups is shown in Figure 2—see the figure note for a detailed explanation.

In the top left grid of this figure, we can see that without any inducement, the fraction of subjects choosing the bottom-right position is only 8%, about half of what would be expected (1/6th or 16.7%—for a two-sided t-test: $p < .03$). Having imbalanced progress bars pushes workers to select the non-focal image. However, offering bonuses has a larger effect. The largest effect by a wide margin occurs when we combine both progress bars and monetary incentives, as shown in the bottom panel.

Workers exhibit a "default bias" and prefer tasks located in more focal positions

As hypothesized, our workers exhibit a strong preference for the image in top-left, focal position and select it 36.4% of the time in our baseline group, more than twice as much as would be expected by random ($p < .003$). The bottom-right (non-focal) position is selected only 8.5% of the time (or less than half as much as would be expected). Figure 2 shows the raw data for the baseline and all other treatments.

In addition to a preference for the focal position, workers also systematically preferred certain *images* over others. A formal test of whether all six images were equally popular is rejected using a chi-square goodness of fit test ($p < 0.014$).[5]

[5]Incidently, the most attractive image for workers was one

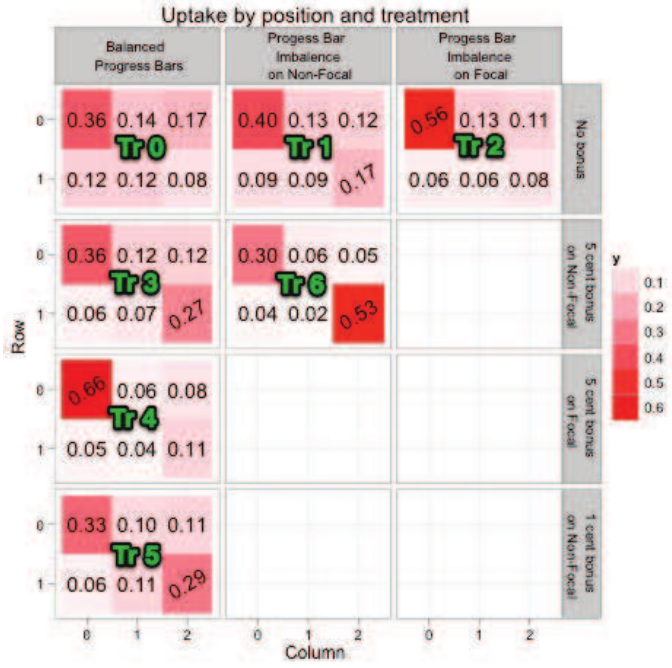


Figure 2: Fraction of subjects selecting different positions within each treatment

Notes: The above matrix includes seven different 2 x 3 grids, each of which represent the raw data showing the proportion of workers who selected each position for the seven treatments. The baseline group, Treatment 0, is shown at the top-left and illustrates the position bias that occurs when progress bars are balanced and without any bonuses. The other 2 x 3 grids are arranged according to the types of incentives employed. Within each 2 x 3 grid, if a number is oriented at 45-degrees, that particular position was incentivized as part of a treatment.

Saliency makes all incentives more effective

We tested the hypothesis of whether more salient incentives would have stronger effects. To evaluate monetary incentives, we either paid a 5-cent bonus in the non-focal position (Treatment 3) or the focal position (Treatment 4). Similarly, the progress bars indicated that the worker choose the non-focal position (Treatment 1) or the focal position (Treatment 2).

At baseline, 8.5% choose the non-focal position and 36.4% choose the focal position. For each position, we thus measure the effectiveness of a treatment by how much it increased the percentage relative to these bases. Our progress bar incentives increased the absolute percentages by 8.6% ($p < .03$) and 19.2% ($p < .002$) for the non-focal and focal positions, respectively; the monetary incentives increased the absolute percentages by 18.7% ($p < .0002$) and 30.0% ($p < .00001$) for the non-focal and focal positions.

which showed toy horses, the least popular image also had the lowest number of labels, suggesting that it may have been avoided due to its difficulty.

Figure 3 shows the size of the change relative to baseline for incentives depending on their type and saliency.

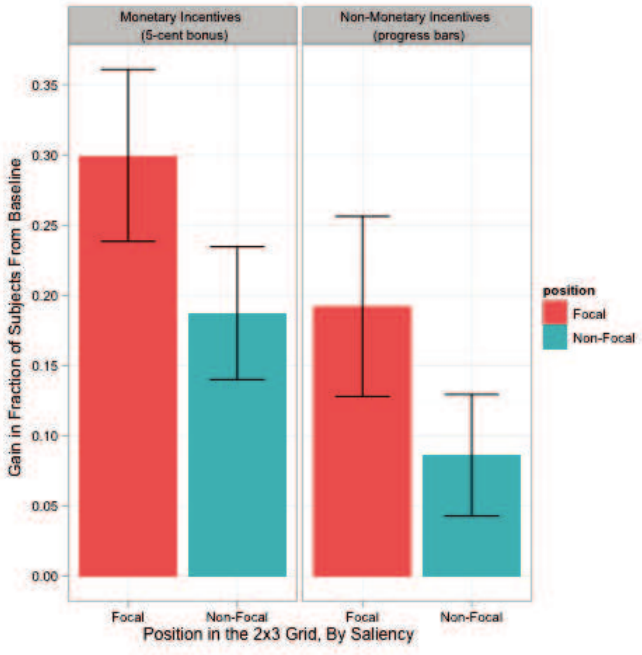


Figure 3: Comparison of treatment effects by saliency of position

Notes: This chart shows the increase in the fraction of people who selected each position relative to the baseline. The data are broken down by whether the incentive was placed in a focal or non-focal position. For example, the first bar indicates that an additional 30 percentage points of people chose the focal position when it had a 5-cent bonus. The above bars correspond to the treatment effects under Treatment 4, 3, 2 and 1 relative to Treatment 0, the baseline. Error bars report standard errors and show that more salient incentives had a significantly greater effect.

Hence for both monetary and non-monetary incentives, we see that more salient positioning substantially increases its effectiveness. This is illustrated in Figure 3.

Progress bars weakly influence whether a worker selects an image

Using data from Treatment groups 0, 1 and 2, we tested whether the arrangement of progress bars causes workers to change which position they select. Surprisingly, the progress bar arrangements can be used as an effective way to influence how workers choose tasks (as shown in the bottom row of Figure 2). Using progress bars in the non-focal position increases the proportion of people who select it to 17.1%, about the level that would be expected if people did not have any bias or preference over positions.

Progress bars have a greater effect on which position a worker selects when placed in a salient, focal position. In the baseline, 36.0% of workers chose the focal image; when the focal image was incentivized using progress bars, this rose to 55.7%. For the non-focal position, the baseline was

8.5% and increased to 17.1% when incentivized. Figure 3 shows same information graphically.

Monetary bonuses strongly influence whether a worker selects an image

We hypothosized that by offering a 5-cent bonus to label the images at certain positions, we could induce workers to label those images. Further, we hypothesized that bonuses would be more effective when placed at the focal position as compared with the non-focal position simply because focal bonuses were more likely to be noticed.

Using data from Treatment groups 0, 3 and 4, we found that bonuses were highly effective, regardless of where they were offered. When the bonus was offered at the focal position, selection of the image at that position increased to 66.4% from the baseline level of 36.4%. When the bonus was offered at the non-focal position, uptake increased from 8.5% to 27.2%.

When monetary bonuses were combined with the progress bars in the non-focal position, the percentage of people choosing the bottom-right image increased from 8.5% to 53.1% for a total of almost 45%. Independently, progress bars raised the percentage almost 9% and bonuses by about 18%. Combining both increased the proportion that selected the non-focal position by an additional 18%.

Don't overpay—small bonuses worked as well as larger ones

We hypothesized that workers would choose the image that was in the position of the highest bonus. Our conjecture is supported by the data, but one caveat for would-be crowdsourcers is that our 5 cent bonus was unnecessary—1-cent was suffice to shift behavior, with no discernible difference in the 1-cent versus the 5-cent treatments.

Using data from treatment groups 0, 3, and 5, we varied the level of the bonus in the non-focal position and to determine whether paying 1 cent versus 5 cents affects the likelihood that workers will choose to work on the image in the non-focal position. As previously discussed, paying a 5-cent bonus in the non-focal position raised the proportion of people selecting that position to 27% relative to a baseline of 8%. When we pay a small bonus of 1 cent, 29% of workers are induced to select the non-focal position. Paying either a 1- or 5-cent bonus more than triples the proportion who select the non-focal position and there is essentially no difference between paying a 1-cent or 5-cent bonus.[6]

Conclusion

As crowdsourcing and human computation systems evolve and expand, the magnitude and efficacy of offered incentives will be of first-order importance: there can be no human computation without the humans, and humans will need to be given the proper incentives.

Our paper shows that incentives and signals—regardless of their type—need to be known and salient in order to be effective. This result is perhaps not too surprising: we would hardly expect workers to be motivated by bonuses they are unaware of or for community members to strive to earn badges that others do not recognize. However, our results have subtler and hence more easily overlooked implications—namely that combining incentives of different types can be very effective. In our experiment, the strongest effects were obtained by using hybrid incentives that combined non-monetary and monetary incentives or by placing monetary incentives in a salient position.

Although monetary incentives worked better than non-monetary incentives in general, a salient, well-positioned non-monetary incentive was as effective as a more costly monetary incentive that was poorly positioned. This suggests that task designers should recognize the role that saliency plays in enhancing or diminishing the power of incentives.

[6]The difference between 27% and 29% is not significant. However, there is a possibility that workers perceived the 1 cent bonus as larger given that it was positioned near a zero-cent bonuses. One way to test this hypothesis would be to introduce another treatment where Treatment 3 and 4 is modified to include all 1-cent bonuses and a single 2-cent bonus (rather than a 5-cent bonus).

Acknowledgments

Thanks to the the Christakis Lab at Harvard Medical School and the NSF-IGERT Multidisciplinary Program in Inequality & Social Policy for generous financial support (Grant No. 0333403). Thanks to Lydia Chilton, Pat DeJarnette, Bruno Ferman, Adam Kapelner, Greg Little, Aaron Shaw, Daan Struyven, Robin Yerkes Horton and seminar participants at the Harvard Labor Economics seminar and for very helpful comments. Dana Chandler acknowledges support from the National Science Foundation in the form of a Graduate Research Fellowship. All plots were made using `ggplot2` (Wickham 2008). Thanks to John Comeau for excellent programming.

References

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

Chilton, L.; Horton, J.; Miller, R.; and Azenkot, S. 2010. Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 1–9. ACM.

DellaVigna, S. 2009. Psychology and economics: Evidence from the field. *Journal of Economic Literature* 47(2):315–372.

Finkelstein, A. 2009. E-ZTAX: Tax Salience and Tax Rates*. *Quarterly Journal of Economics* 124(3):969–1010.

Gneezy, U., and List, J. 2006. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica* 1365–1384.

Goldberg, J.; Stimson, M.; Lewenstein, M.; Scott, N.; and Wichansky, A. 2002. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, number 650, 51–58. ACM.

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

Hornof, A. 2004. Cognitive Strategies for the Visual Search of Hierarchical Computer Displays. *Human-Computer Interaction* 19(3):183–223.

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

Horton, J. J.; Rand, D.; and Zeckhauser, R. J. forthcoming. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*.

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

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

Pallais, A. 2010. Inefficient Hiring in Entry-Level Labor Markets. *MIT mimeo*.

Shaw, A.; Horton, J.; and Chen, D. 2010. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*.

Simon, H. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*.

Thaler, R., and Sunstein, C. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. Yale Univ Press.

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

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

Designing Incentives for Inexpert Human Raters

Aaron D. Shaw
UC Berkeley
410 Barrows Hall #1980
Berkeley, CA 94720
adshaw@berkeley.edu

John J. Horton
Harvard University
383 Pforzheimer MC
56 Linnaean St
Cambridge, MA 02138
horton@fas.harvard.edu

Daniel L. Chen
Duke University
210 Science Dr
Office 3020
Durham, NC 27708
dchen@law.duke.edu

ABSTRACT
The emergence of online labor markets makes it far easier to use individual human raters to evaluate materials for data collection and analysis in the social sciences. In this paper, we report the results of an experiment — conducted in an online labor market — that measured the effectiveness of a collection of social and financial incentive schemes for motivating workers to conduct a qualitative, content analysis task. Overall, workers performed better than chance, but results varied considerably depending on task difficulty. We find that treatment conditions which asked workers to prospectively think about the responses of their peers — when combined with financial incentives — produced more accurate performance. Other treatments generally had weak effects on quality. Workers in India performed significantly worse than US workers, regardless of treatment group.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Interaction Styles; H.3.3 Information Storage and Retrieval: Information Search and Retrieval; J.4 Social and Behavioral Sciences: Economics, Sociology

General Terms
Economics, Sociology, Experimentation, Measurement, Human Factors

Author Keywords
Experimentation, Amazon Mechanical Turk, Human Computation, Crowdsourcing, Search, Content Analysis

BRIEF SUMMARY
We compare incentive schemes in an online labor market experiment and find that asking subjects to consider the answers of their peers produces more accurate performance on a content analysis task. Workers in India and workers with lower web-browsing skills also performed worse than their peers on the task.

CSCW 2011, March 19-23, 2011, Hangzhou, China.

INTRODUCTION
The binding constraint in much observational, empirical research in the social sciences is finding data with useful, well-measured indepedent and dependent variables. Often, compelling research questions require the quantification of complex constructs such as trustworthiness, beauty, or aggression. Since these kinds of measures are unlikely to appear in observational data sets, researchers must look at primary source material and then classify it according to some coding scheme. A recent economic study asked subjects to assess the trustworthiness of loan-seekers based on their photographs on the social lending site Prosper.com and then used these ratings to predict loan outcomes[10]. Another example used amateur evaluations of short debate clips from gubernatorial elections and found that these evaluations were predictive[3]. Often times these qualitative coding tasks require human judgment, but not any expertise. While this makes them ideal for inexpert raters, the tasks themselves are often tedious and time-consuming, and finding research assistance to perform them may be difficult or expensive.

An emerging phenomenon — the online labor market — can scale the process of qualitative coding (also known in some social science circles as "content analysis") using large numbers of non-experts. Previous research has discussed the potential advantages of online labor markets in terms of the cost, scale, and validity of experimental data collection.[16] However, these studies have not addressed the opportunities and constraints of applying distributed labor to content analysis tasks. In particular, while the number of workers participating in online labor markets makes it relatively easy to attract many judgments for any task, it is difficult to elicit and synthesize high-quality judgments from non-expert raters collaborating remotely. Among the foremost practical challenges of this kind, the design of optimal incentives schemes to facilitate this peculiar form of cooperative work has received scant scholarly attention. Prior economic, sociological and psychological research offers much theoretical guidance, but little empirical evidence as to the sorts of incentives that elicit the highest quality judgments from non-expert raters.

In this paper, we present the results of a controlled experiment that directly compares the effects of fourteen different incentive schemes within the context of an online labor market. The incentive schemes encompass a wide variety of existing research into human cooperation, labor, motivation and behavior. We test the incentives using a single,

non-expert content analysis task, for which we obtained validated answers prior to administering the experiment. We then compare the aggregate performance of workers in the different treatment conditions in order to determine which incentive schemes elicit the most accurate judgments in comparison to the control condition.

Use of Online Labor Markets

In online labor markets, workers from around the world perform data processing tasks for money. While some sites focus on skilled work like computer programming (e.g., oDesk, Elance, Guru), Amazon's Mechanical Turk (MTurk) is intended for small, simple and discrete tasks and thus is probably the most directly useful for researchers. The challenge of tapping this resource is that raters are inexpert and there is sometimes a high degree of inter-rater disagreement, regardless of the measure. The low cost of raters make large numbers of ratings possible, but this volume of data also prohibits a hand-curated approach to selecting high-quality raters.[1]

Several papers in the computer science literature have used online labor markets such as MTurk to conduct experiments [21, 29, 28]. Horton, Rand and Zeckhauser discuss the social science potential of online experiments in these markets, focusing on how challenges to validity can be overcome [16]. There already exists a small literature on crowdsourcing from a social science perspective [18, 24, 15, 7, 6]. New tools are also being developed that make experimentation easier [23].

Obtaining Quality Work

In online labor markets, the usual rules of labor supply generally apply: more money attracts more workers on both the extensive margin (i.e., more workers are willing to participate at all) and the intensive margin (i.e., workers that participate work longer or produce more). However, attracting more workers does not necessarily lead to better work. While earlier work [30, 19, 29, 14, 9] has focused on techniques for filtering and processing judgments of inexpert human raters, we focus on how to produce better judgments in the first place.

Some work has already been done in this vein. A recent experimental paper by Chandler and Kapelner [5], conducted in MTurk, looked at how knowledge about the purpose of a task affected quality and labor supply. US-based subjects who knew they were labeling cancer cells in an image produced more output than those who did not. Interestingly, they found no evidence of similar effects for non-US workers. The same authors also recently conducted an experiment in which they demonstrated that slowing down the presentation of survey questions increased comprehension [20].

Content Analysis

In some kinds of research, human judgments can be evaluated against objective, correct answers. This is the case for tasks such as image labeling or character recognition, where accurate automated techniques remain costly or unavailable. In others, human judgments are important precisely because they incorporate subjective perceptions, which may be central to the topic of study. This is the case for many types of content analysis tasks, where researchers aim to identify certain qualities or patterns in textual materials that evade automated detection. In both objective and subjective variants, the challenge of developing techniques to aggregate individual judgments as well as to assess their precision and accuracy has given rise to several different methodological techniques, some of which we review as background to the method we used in this study.

Useful methodological approaches to this type of problem have emerged among scholars conducting content analysis of textual materials. Until recently, content analysis techniques have relied on multiple researchers implementing a qualitative labeling or coding scheme of the same text(s), and then using specifically adapted correlation statistics to evaluate inter-rater (or intercoder) reliability [22, 8]. The primary advantage of these approaches lies in the ability to measure empirically the reliability of seemingly subjective observations. The cost of such precision, however, is often quite high in terms of time and labor, making such analysis prohibitively expensive when the scale of data collection and analysis grows large. Recent work by Hopkins and King has demonstrated that machine-learning tools and techniques can overcome these limitations while retaining high confidence in the precision and accuracy of results [14].

Our Approach

A variety of papers across the social sciences have studied human motivation. This literature is far too voluminous to summarize here; much of it is also captured by folk wisdom or even in management cliches. What is certainly not known is the relative merits of different motivations and how they apply in online contexts. For example, does offering workers more money improve effort and hence quality? This lack of knowledge motivated this study, in which we created a large number of treatment groups and recruited a vast number of subjects. While this "kitchen sink" approach creates some problems of analysis, it does afford our observations greater breadth of comparison. We review the different motivational frameworks in greater depth below.

Our Task

For our task, we asked subjects recruited from MTurk ("Turkers") to complete six closed-ended, qualitative content analysis questions using an online survey interface. All subjects in all treatment groups (except one of the control groups, which only answered demographic questions) were directed to analyze the Kiva.org website and then presented with the same six questions in the same order and with the same answer choices. The questions asked subjects to conduct content analysis similar to that used in an earlier study by [4] to assess US political blogs. For any questions, workers could choose to leave a blank response.

[1] Several innovative start-up companies, such as Crowdflower are offering services as intermediaries. Clients bring them tasks amenable to the crowdsourcing approach and they break the tasks down, recruit workers and ensure quality results.

Overview of Results

Our results varied by question as well as by treatment condition. On the two easiest questions, the Turkers uniformly performed much better than random guessing and only a couple of the treatments seemed to produce any (small) effect at all. By contrast, the results for the three difficult questions varied more widely. In one case, the Turkers' performance was much worse than chance. At the same time, the variance in responses to these questions also revealed stronger treatment effects. Aggregating the results from each condition across all five questions, the Turkers performed better than chance. More importantly, a few treatments proved to be markedly more effective than the others, producing significant improvements in average answer quality when compared against the control condition. We discuss the experimental design, data collection and results in greater depth below.

METHODS AND MATERIALS

Content Analysis Task

In order to establish a reliable standard against which to judge the performance of the workers, we also administered the same questions about the same website through an identical web interface to a group of five research assistants prior to conducting the experiment. On all of the questions included in the study, at least four of the five research assistants gave identical responses, suggesting a high degree of intercoder reliability. Independent of the research assistants, one of the authors also collected his own answers to the questions, agreeing with the prevailing answer provided by the research assistants in every case. We used these responses as validated (i.e., gold standard) answers to each question.

The first two questions followed a multiple choice format, in which subjects were asked to identify whether (1) a privacy policy; and (2) "avatars" or other visual representations of user identities were present on the site. For both of these questions an "uncertain" answer choice was also available. The third and fourth questions asked subjects to assess how frequently members of the site engaged in specific behaviors (ranking or rating (3) content and (4) other users) using a five point scale ranging from "Very frequently" to "Very rarely or never." Finally, the last two questions asked subjects to identify whether specific features related to (5) social networking and (6) revenue creation were present or not on the site. In these last two, subjects could check boxes to select any combination of answer choices from a pre-defined list.

The first of the six questions (about whether or not the site had a privacy policy) was presented prior to treatment. We report the results for this pre-treatment question but do not include it as part of our outcome performance measurement. A copy of the questions as they appeared in the experimental interface is available on Horton's website.[2]

The dependent variable of our study was the number of correct answers to the five post-treatment information-seeking questions per subject.[3] We considered blank responses incorrect answers for all questions. After coding responses to identify which ones each subject answered correctly (i.e., in agreement with the gold standard response), we aggregated the number of correct answers per subject. The outcome measure is therefore an integer (count) with a value between zero and five. As we describe in further detail below, the subjects recruited through MTurk performed better than chance - estimated as random guessing between all available answer choices for every question - on four of the five post-treatment questions.

The demographic questions asked subjects to provide their age; gender; country of residence; education level; language skills; employment status; household size; and internet skills. We included them to increase precision in our treatment estimates as well as to verify that our randomization was valid (we discuss the rationale for this choice in further detail below).

Conduct of the experiment

Recruitment was conducted through the MTurk online labor market, where we advertised a brief information-seeking task. Recruitment materials included a description of the study as well as a set of example questions, all of which were included in the actual job, but none of which were among the post-treatment questions included in our outcome variables of interest. Subjects were not informed that they were participating in a study at the time of recruitment so as to preserve the "natural" environment of the field experiment in the online labor market. In the task description, we explained that workers would be paid $0.30 for completing the task. Given the length of the assignment and the fact that workers could only complete our job once (many jobs on MTurk allow workers to return multiple times), this payment rate was comparable with many other jobs posted to the MTurk marketplace.

Upon agreeing to accept the task on the MTurk website, subjects were instructed to click a hyperlink pointing to a private server at an anonymized URL. While we were not able to collect data on how many individuals saw our recruitment materials, once a worker accepted our task, their unique MTurk user ID was assigned randomly to one of the treatment or control conditions and (together with their IP address and the information about treatment assignment) stored by a database on our server. As a result, we were able to use these different pieces of stored identifying information to block individual subjects from completing the study more than once or from being exposed to more than one of the experimental manipulations. While there is some possibility that individuals could possess more than one account on the MTurk platform and thereby might have circumvented these protections, such behavior is expressly prohibited by

[2] http://goo.gl/9CVa5

[3] In the case of the checkbox questions – numbers (5) and (6) – we coded any response including the gold standard answer as correct. Obviously, in the case of a question where we did not know the correct answer ahead of time, a much different process would be needed to identify the best response. As such filtering processes were not the focus of this study, we refer consideration of this topic to the work of others.[29]

the site's terms of service and Amazon actively polices violations (indeed, one of the authors of the study had the somewhat embarassing experience of losing his MTurk account as a result of attempting to create multiple user names in order to test a pilot version of an earlier study). Furthermore, the payoff for circumventing the system protections on our job (which required a little more than 2000 unique judgments) were very low in comparison with some of the large scale jobs on the site which frequently elicit hundreds of thousands or even millions of individual judgments. As a result, we feel confident in the integrity of both the randomization as well as the different treatment conditions.

Once Turkers clicked through to our server, the experimental instrument was administered through a web-based survey interface. Subjects were presented with a single page containing the version of the instrument corresponding to their treatment assignment. Each version of the instrument began with some general instructions about the task, and (in all conditions except for the demographic control) a link to the URL of the site that would serve as the topic of the questions (Kiva.org). These were followed by several pre-treatment questions about the site. Then, we introduced the experimental manipulations (usually consisting of a block of text) followed by the post-treatment questions and any treatment-specific materials. Finally, the instruments concluded with a series of demographic questions.

Overview of Treatments: Social, Financial, and Hybrid Incentives

The experimental manipulations we introduced consisted of framing the information-seeking questions in distinct ways using a series of "social" and "financial" incentives. Together, these incentive schemes encompass a number of salient theories of human motivation drawn from several social sciences. Generally, the social incentives emphasized non-monetary rewards or punishments for performing our task whereas the financial incentives offered monetary rewards (bonus payments) for good performance or punishments (lost bonus payments). Some frameworks were hybrids that combined social and financial incentives. In total, we tested fourteen different incentive frameworks and compared subject performance in each condition against a control condition that involved no framing incentives beyond the baseline compensation offered for completing the job. We also included a second control group in which subjects responded only to the pre-treatment and demographic questions used in the other conditions. All subjects who completed the task were given the baseline compensation. Because of some technical complications, we paid all subjects the largest amount they could have received from their experimental treatment in order to avoid under-paying any deserving subjects.

We describe all control and treatment conditions in further detail below. For each of the conditions (listed in bold) we note in parentheses whether it is social, financial or hybrid and include the full treatment text. Where appropriate, we also include references to relevant studies in which comparable incentives were found to effect behavioral outcomes.

Control Conditions

Control Workers were presented with all pre-treatment, post-treatment and demographic questions.

Demographic Workers were presented with pre-treatment and demographic questions only.[4]

Treatment Conditions

Tournament scoring (social) "For some of the following five questions, you will be in competition against another worker. After this HIT is completed, we will compare your accuracy on these questions against the accuracy of another worker who we will select at random. We will report the results of the competition to you when we process your payment."

Cheap Talk — Surveillance (social) "After this HIT has been completed, your answers to these questions will be reviewed for accuracy."

Cheap Talk — Normative (social) "It is your job to provide accurate answers to these question. It is important that you do your job well."

Solidarity (hybrid) "For some of the following five questions, you have been assigned to the Red team. You and your teammates have the opportunity to earn bonuses based on your collective performance. After the HIT has been completed, we will verify the answers that you all submitted for these questions (independent of the website you are analyzing) and compare your team's performance with another group of workers completing this HIT. If your team wins, you will all receive a bonus."

Humanization (social) "Before you complete the questions, I just wanted to thank you again for doing this work. My name is Aaron."[5]

Trust (social) "Thank you for completing the first set of questions. Here is your confirmation code, which you may paste into the field on the original HIT page at any time to receive payment. We trust that you will still complete the questions below to the best of your ability. Your confirmation code and payment for this HIT will not change based on the answers you submit."[6]

Normative priming questions (social) "Before answering the next set of questions about the website, we want to

[4]Whenever possible, the demographic questions were taken verbatim from the 2005 codebook of the World Values Survey [1]. As we described later in the paper, we also borrowed two questions about Internet-use skills from Eszter Hargittai [13].

[5]This treatment text was accompanied by a photo of one of the authors.

[6]In order to make this treatment condition consistent with the design of all other conditions, all workers were asked to submit a completion code when they finished the job. In every condition except this one, we provided these completion codes once the task had been finished and the answers to all questions submitted to our server. Compensation was not conditional on submitting the completion code in any of the conditions.

ask you a few questions about yourself and your attitudes about work."[7]

Reward Accuracy (financial) "After this HIT has been completed, we will verify the correct answers for at least one of the following five questions. For each 'trap door' question we will increase your total pay by 10% if you answered it correctly. You will not receive this bonus if you do not answer the 'trap door' question(s) correctly."

Reward Agreement (financial) "After this HIT has been completed, we will review the answers for at least one of the following five questions. For each of the questions we review, we will reward you for agreeing with the answers provided by the majority of other workers who complete this HIT. The reward will be a bonus of 10% for every agreement."

Punishment Accuracy (financial) "After this HIT has been completed, we will verify the correct answers for at least one of the following five questions. For each one of these 'trap door' questions we will penalize you 10% of the bonus that you would have received if you answered it incorrectly."

Punishment Agreement (financial) "After this HIT has been completed, we will review the answers for at least one of the following five questions. For each of the questions we review, we will penalize you if you disagree with the majority of other workers who complete this HIT. The penalty will be a deduction of 10% from the total bonus you could have earned if your answer had agreed with the majority."

Promise of Future Work (financial) "After this HIT has been completed, we will review the performance of each worker on the following five questions. If you perform better than average, you will have the opportunity to work on future jobs with us."

Bayesian Truth Serum or BTS (financial) "For the following five questions, we will also ask you to predict the responses of other workers who complete this task. There is no incentive to misreport what you truly believe to be your answers as well as others' answers. You will have a higher probability of winning a lottery (bonus payment) if you submit answers that are more surprisingly common than collectively predicted."[8]

[7]This text was followed by a series of questions drawn from the General Social Survey inquiring about subjects' agreement with statements indicating positive attitudes towards responsibilities and hard work. The statements, in order, were "People who don't work become lazy"; "Work is a duty toward society"; "Work should always come first, even if it means less free time"; "Work is a person's most important activity"; "I see myself as someone who does a thorough job."

[8]The design for this treatment comes from [25] who used a near identical method in an effort to elicit honest opinions from their research subjects. After data collection, the responses were subsequently weighted based on the aggregate predicted distributions of the respondents. For our own purposes, we were merely interested in the question of whether presenting our task in a similar way would have a meaningful effect on qualitative information seeking. The results we present do not involve any of the weighting procedures used by Prelec. We refer interested readers to the original paper for more detailed information about this technique.

Betting on Results (financial) "For the following five questions, you will have the opportunity to win bonuses. After completing the questions, we will let you bet a portion of your payment on the accuracy of your responses."

Data Collection

The experiment ran from June 2 through September 23, 2009. During that time, we collected a total of 2159 unique subjects, of whom 2055 completed the study and 104 dropped out after treatment assignment. Because we used a random treatment assignment function (instead of stratified random assignment), the distribution of subjects across conditions was unequal, ranging between 113 and 167 subjects per condition. Applying Pearson's χ^2 test to a contingency table with the counts of attriters and compliers across all of the treatment and control groups suggests that attrition was not significantly different from random ($p = 0.919$).

We also ran a regression of all the demographic covariates against treatment condition to test whether our randomization worked. The model was not significant and none of the variables had a significant association with treatment assignment. As a result, we conclude that randomization was successful.

Following the completion of data collection, we discovered that database storing our records from the study had stored inaccurate values for three of the subjects. As a result, we excluded the results from these three subjects from all subsequent analysis, with the exception of the calculation of the total number of subjects assigned to each treatment group used to generate our estimates of treatment effects (see below).

Statistical Analysis

In all of our estimates of treatment effects, we correct for the increased probability of Type 1 errors when conducting multiple hypothesis tests in an experiment with many treatments by using the single-step Bonferroni correction to adjust our p-values [27, 17]. This correction has the advantage of simplicity as well as strong control of the Familywise Error Rate (FWER) in a context where the comparisons being tested are unordered [26].[9]

We used Intention-To-Treat (ITT) estimators to calculate the average effect of each treatment compared against the control condition. What this means practically is that subjects that quit after assignment to a group were still included in calculations as answering incorrectly. ITT estimators have the advantage of correcting for potentially confounding effects of attrition and avoiding the bias introduced into the analysis of many randomized experimental results by regression estimates [12, 11].

RESULTS

Performance on Individual Questions

Looking at the percentage of correct responses per question across all conditions (except demographic control), subject performance varied significantly from chance (random

[9]We calculate these corrections using the "multtest" package in R.

guessing among the available answer choices) for all five questions (see Table 1).[10] On four of the five, subjects performed better than chance, whereas the question about revenue streams elicited performance that was significantly worse than chance.

Table 1. Performance on Individual Questions (All Conditions)

	Actual % correct	Predicted % correct (random guessing)
Avatars	73.2	25
Content rank/rate	25.6	20
User rank/rate	28.7	20
Revenue streams	47.6	50
Soc. network features	62.8	50

χ^2 test indicates all differences significant ($p \leq 0.05$)

Comparing the percentage of correct answers across questions and across experimental conditions reveals fairly consistent performance from each treatment group despite the substantial variation across questions (see Figure 1).[11]

Aggregate Performance (All Five Questions)

Figure 2 illustrates aggregated worker performance across all five questions and all experimental conditions. On average, subjects did significantly better than chance, which would have yielded a mean of approximately 1.58 questions correct. The actual distribution of responses is strikingly close to normal, with a slight concentration at 2 and a mean of 2.38.[12]

The results of our ITT estimation of average treatment effects (ATE) are reported in Table 2.[13] To facilitate the readability of the table, we order all treatment conditions by the absolute size of their estimated effects and only report $p \leq 0.05$.

As described above, we used the "simple" Bonferroni correction for the difference of means comparisons between each treatment group and the control condition. The results suggest that only two of our treatments produced a significant improvement in worker performance over the control:

[10] We used χ^2 tests for goodness of fit to calculate these comparisons between the distribution of correct responses and predicted probabilities of producing correct answers through random guessing for each question.

[11] We did not conduct hypothesis tests comparing average treatment effects for each question. Such question-level effects were not our primary outcome variables in part because of the specificity of the content of each question and the fact that we looked at responses only from a single website. See the Discussion section below for additional consideration of this topic.

[12] This mean reflects only the performance of compliers - not the full set of subjects exposed to treatment. This corrected (ITT) sample mean was 2.26.

[13] The ITT estimate of the ATE captures the mean difference in aggregated performance between the subjects in each treatment condition and the subjects in the control group. The estimates themselves are identical with the results of a linear regression on the same data. The standard errors are different as are the underlying p-values[12, 11]. As discussed above, all p-values have been corrected using the simple Bonferroni correction procedure [27, 17].

Figure 1. Performance Distributions - All Conditions

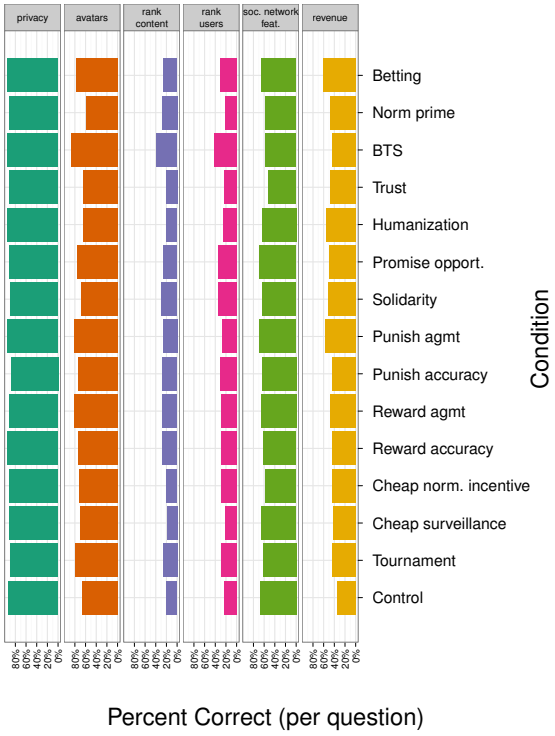


Figure 2. Performance Distribution - All Conditions

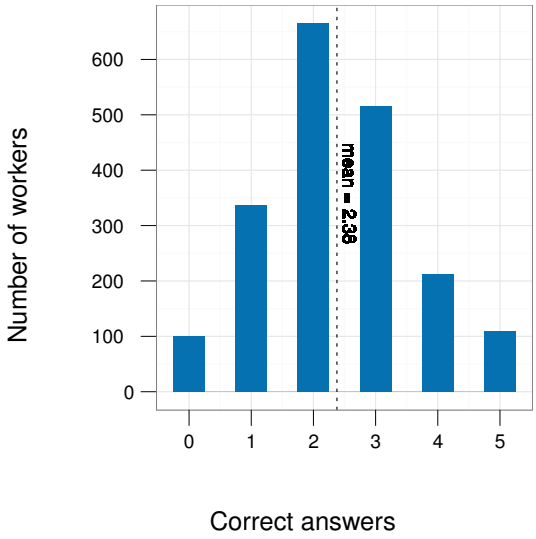


Table 2. Average Treatment Effects (ATE) on Aggregate Performance

	Mean	ATE†	Std. Err.	p-val.‡
Control	2.079	NA	NA	NA
BTS	**2.549**	**0.471**	**0.132**	**0.017**
Punish-agmt.	**2.538**	**0.459**	**0.131**	**0.015**
Betting	2.438	0.359	0.137	
Reward-agreement	2.421	0.342	0.135	
Promise-opportunity	2.404	0.326	0.138	
Tournament scoring	2.310	0.232	0.142	
Solidarity	2.296	0.217	0.149	
Punish-accuracy	2.275	0.197	0.131	
Reward-accuracy	2.214	0.136	0.139	
Humanization	2.171	0.092	0.142	
Trust	2.029	-0.050	0.137	
Cheap talk-surveil.	2.027	-0.052	0.131	
Normative Priming	2.057	-0.021	0.142	
Cheap talk-norm.	2.075	-0.003	0.141	

† ATE calculated using Intention-to-Treat (ITT) estimators.
‡ p-values reported ≤ 0.05.

Punishment-agreement and Bayesian Truth Serum. In each case, the effect was approximately .5 above the mean outcome in control (2.08). Both were significant at $p \leq 0.05$.

Post-hoc Demographic Analysis

To evaluate whether any demographic factors may have affected our estimates, we ran an ordinary least squares (OLS) model on the dependent variable (aggregate performance, or score out of five), incorporating the full set of demographic control variables together with the treatment assignments. The results of this "full" model (not reported here) suggested that three covariates may have had a significant association with subject-performance despite the randomization: web-use skills[14], household size, and country of residence. To zero-in on any potentially confounding effects of these variables, we ran a second model that included only the outcome, the treatment conditions, and these three covariates.[15]

The second model (reported in Table 3) suggests a significant, negative association between performance on our outcome measure, poor web skills and residence in India (both covariates were significant at the $p \leq 0.001$ level after correcting for multiple comparisons). Remarkably, the point estimate of the association between residence in India and the outcome variable dwarfed any of our estimated treatment effects. Again, treatment conditions and covariates are sorted by point estimate size to facilitate readability.

Table 3. OLS Regression on Aggregate Performance

	Estimate	Std. Err.	p-value†
(Intercept)	1.851	0.153	0.000
India resident	-0.739	0.068	0.000
BTS	0.596	0.138	0.000
Punishment-agreement	0.482	0.137	0.008
Betting	0.437	0.139	0.031
Promise-opportunity	0.398	0.139	
Tournament scoring	0.358	0.143	
Reward-agreement	0.310	0.139	
Solidarity	0.291	0.144	
Reward-accuracy	0.232	0.139	
Punishment-accuracy	0.230	0.133	
Web skill	0.147	0.024	0.000
Humanization	0.136	0.141	
Cheap talk-normative	0.131	0.140	
Household size	-0.048	0.018	
Trust	0.047	0.138	
Normative Priming	0.039	0.138	
Cheap talk-surveillance	0.035	0.147	

Adjusted $R^2 = 0.127$
† p-values reported ≤ 0.05.

DISCUSSION

Our results suggest a significant, positive effect of two treatment conditions - Punishment for disagreement with other subjects, and "Bayesian Truth Serum" (BTS) - on worker performance in a qualitative content analysis task on MTurk. Several of the other "financial" incentive schemes produced large point estimates of treatment effects, but were not significantly different from the control condition. None of the purely "social" incentive schemes altered performance significantly. This suggests that workers in the MTurk environment may not respond to these sorts of motivational levers.

Even though the two most effective conditions – BTS and Punishment for disageement – are both examples of financial incentive schemes, the fact that they alone succeeded does not imply a ringing endorsement of monetary incentives over social incentives by the workers on Mturk. Rather, the challenge of these results lies in explaining why these *particular* financial incentive schemes appeared to work where so many others did not.[16]

We contend that the most likely explanation of these results hinges on the fact that both the BTS and Punishment-disagreement conditions tied worker payoffs to their ability

[14]To measure this variable, we borrowed a survey item from an instrument designed, validated, and implemented by Eszter Hargittai in several of her studies.[13] The item asks subjects about their understanding of two web-browsing tools: "tabs" in an internet browser and RSS feeds. Hargittai found that both items correlate highly with independent measures of web-browsing and Internet skill.

[15]The fact that country of origin was significant suggested a result consistent with previous findings about the differences between workers from India and the US[19]. As a result, we re-coded country of residence as a binary variable, indicating whether workers self-reported as residing in India or not.

[16]One of the anonymous CSCW reviewers suggested comparing the financial incentives versus the social incentives in another way by grouping the conditions into clusters and estimating treatment effects between the different clusters. While our research design supports this line of inquiry, we choose not to pursue it here for two reasons. First, we did not conduct any preliminary testing to validate our classification of the different treatments into one or the other group. Second, our findings suggest that even the financial incentives were not purely financial in any sense (see the rest of this Discussion section for more on this topic).

to prospectively reason about the performance of their peers. However, what specific mechanisms can account for these effects in each condition?

When compared with the other treatments and control conditions, BTS likely had two effects: (a) it created some confusion among subjects about how exactly they were being evaluated; and (b) it created an incentive for subjects to think carefully about the respones of other subjects. The combination of confusion and cognitive demand probably elicited greater engagement with the question, and this engagement in turn probably drove better performance. In the case of BTS, we should underscore that the treatment effect is not due to any manipulation of the responses or the predictions provided by the subjects regarding the distribution of responses. In this regard, we did we did not follow Prelec's original design and use the BTS to adjust or filter subjects' answers.[25] Instead, we simply used it as a contextual manipulation. Given that we did not provide much information about the BTS design to the subjects performing the task, it also seems unlikely that they would have understood the analytical mechanisms proposed by Prelec.

The effect of the Punishment-disagreement condition raises a distinct set of concerns insofar as it closely resembles the Reward-agreement condition. In theory, both conditions ask workers to perform a similar set of calculations about the likely responses of other Mturk workers. However, the key difference between the two stems from the role of punishment and reward in the context of relational contracts in online labor markets. In Mturk, punishment of workers by requesters is consequential in a way that rewards is not: workers can be banned from the site if their work is rejected by requesters. As a result, even though we did not claim we would block any worker as part of the Punishment-disagreement condition, by demonstrating our willingness to punish we may have inadvertently suggested that there could be consequential results (like rejection) to poor performance on that particular question. In contrast, the language of rewards and bonus payments used in the Reward-agreement condition would not carry any of these conotations.

It is noteworthy that although the Reward-agreement condition did not have significant effects, it did produce one of the larger point-estimates, suggesting that prospective reasoning by subjects about their peers may have played an attenuated role in that group as well. However, the point estimates for the Betting and "Promise of future opportunity" conditions were similar to that for Reward-agreement (and also not significantly different from the control condition). Both of these conditions asked workers to engage in prospective reasoning, but entail completely different mechanisms from the agreement-based treatments.

We also find a strong association between residence in India, web skills, and task performance. This implies that culturally specific knowledge and experience online may mediate workers' ability to perform the sort of qualitative information-seeking task we asked them to do here.[17]

[17]As a comparison across Mturk workers in India and the US was not part of our original research design or hypotheses, we chose not to compare treatment effects across the two populations in a more purposive manner. Nevertheless, our findings here strongly suggest that future research should conduct cross-national comparisons of workers in online labor markets and other settings. Based on our results, we anticipate significant differences of motivation and performance along these lines.[2]

At the same time, we do not believe that these demographic factors undermine our findings with regards to the effects of Punishment for disagreement and Bayesian Truth Serum. While the association between web skills, residence in India and our outcome variable were quite strong, the point estimates for the effects of these two treatment conditions hardly changed and remained significant at the $p \leq 0.01$ level. This suggests that the effects we observed for the treatment conditions (at least the significant ones) were robust and supports our earlier claim the randomization worked as a means for distributing these sub-populations evenly across the different treatment groups.

CONCLUSIONS AND FUTURE WORK

The connection we observed between qualitative information-seeking performance and treatment conditions asking workers to engage in prospective reasoning about their peers merits further analysis in online and offline settings. In addition, future studies conducted online and with international subject-populations should consider the effects of potentially confounding covariates such as country of origin and web-use skills when designing comparable studies. In our case, the randomization proved effective, but this might not be possible in other settings.

Our results suggest that Turkers appear to have a wide range of abilities and that some task framings may elicit higher quality performance than others. It is worth emphasizing that we do not utilize any of the quality-control techniques discussed elsewhere for filtering data generated in Mturk and similar environments[30, 19, 29, 14, 9, 5, 20]. As a result, we do not use our findings as a basis for any general claims about the utility (or lack thereof) of Mturk for crowdsourced content analysis as a reliable method of data collection. Indeed, we hypothesize that incorporating additional quality control techniques on top of the effective treatment conditions reported in this study could amplify quality improvements beyond what we report here and what other studies have found. Subsequent research is needed to determine the effects of interactions between the worker characteristics, motivational framing, and other interface design manipulations.

Finally, we believe that similar studies should be conducted among other populations online where the existing institutional structure favors other motivational criteria. The rules and norms of the MTurk marketplace favor financial incentives, punishment-oriented consequences and arm's-length relational contracting over more personalistic or socially-oriented modes of exchange. Therefore, it would be interesting to know whether the same incentive schemes would work among a population of Wikipedia contributors who are accustomed to performing similar tasks without any financial payoffs at all.

ACKNOWLEDGMENTS

The authors thank Yochai Benkler, the Berkman Center for Internet and Society and the Law Lab for supporting this work. We are especially grateful to Jason Callina and Anita Patel for creating and administering Not-a-Number, the web application that handled the data collection and storage for this study. Thanks also to Eszter Hargittai for giving us permission to use her survey questions. Finally, Dana Chandler as well as five anonymous reviewers gave us thoughtful feedback on an earlier draft.

John Horton thanks the NSF-IGERT Multidisciplinary Program in Inequality & Social Policy for generous financial support (Grant No. 0333403).

Aaron Shaw thanks Jas Sekhon, Erin Hartman, and Daniel Hidalgo for their patience. He also thanks Vaughn Hester and John Le of Crowdflower for their feedback on an earlier draft.

Daniel Chen is Assistant Professor of Law and Economics at Duke University. Work on this project was conducted while he received financial support from the Institute for Humane Studies, John M. Olin Foundation, and the Ewing Marion Kauffman Foundation.

All figures were created with the excellent R package ggplot2, developed by Hadley Wickham [31].

REFERENCES

1. World values survey 1981-2008 official aggregate v.20090901. *World Values Survey Association (www.worldvaluessurvey.org), Aggregate File Producer: ASEP/JDS, Madrid*, 2009.
2. J. Antin and A. Shaw. Social Desirability Bias in Reports of Motivation on Mturk. In *CSCW Horizons Workshop*, 2011.
3. D. J. Benjamin and J. M. Shapiro. Thin-slice forecasts of gubernatorial elections. *Review of Economics and Statistics (Forthcoming)*, 2009.
4. Y. Benkler and A. Shaw. A tale of two blogospheres: Discursive practices on the left and right. *Berkman Center for Internet and Society Working Paper Series*, 2010.
5. D. Chandler and A. Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *University of Chicago mimeo*, 2010.
6. D. L. Chen. Markets and Morality: How Does Competition Affect Moral Judgment? *Working Paper, Duke University School of Law*, 2010.
7. D. L. Chen and J. J. Horton. The wages of paycuts: Evidence from a field experiment. *Working Paper*, 2009.
8. J. Cohen. A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, 20(1):37–46, 1960.
9. J. Downs, M. Holbrook, S. Sheng, and L. Cranor. Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 2399–2402. ACM, 2010.
10. J. Duarte, S. Siegel, and L. A. Young. Trust and credit. *SSRN eLibrary*, 2009.
11. D. A. Freedman. On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, 2(1):176–196, 2008.
12. D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, 23(2):237–249, 2008.
13. E. Hargittai. An update on survey measures of Web-Oriented digital literacy. *Social Science Computer Review*, 27(1):130–137, 2009.
14. D. Hopkins and G. King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010.
15. J. Horton and L. Chilton. The labor economics of paid crowdsourcing. *Proceedings of the ACM Conference on Electronic Commerce 2010*, 2010.
16. J. J. Horton, D. Rand, and R. J. Zeckhauser. The Online Laboratory: Conducting Experiments in a Real Labor Market. *SSRN eLibrary*, 2010.
17. J. C. Hsu. *Multiple comparisons: Theory and Methods*. CRC Press, 1996.
18. B. A. Huberman, D. Romero, and F. Wu. Crowdsourcing, attention and productivity. *Journal of Information Science (in press)*, 2009.
19. P. Ipeirotis. Demographics of mechanical turk. *New York University Working Paper*, 2010.
20. A. Kapelner and D. Chandler. Preventing Satisficing in Online Surveys. In *Proceedings of 2010 CrowdConf*, 2010.
21. A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. *Proceedings of Computer Human Interaction (CHI-2008)*, 2008.
22. K. H. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Inc, 2nd edition, 2003.
23. G. Little, L. B. Chilton, R. Miller, and M. Goldman. Turkit: Tools for iterative tasks on mechanical turk. *Working Paper, MIT*, 2009.
24. W. Mason and D. J. Watts. Financial incentives and the 'performance of crowds'. In *Proc. ACM SIGKDD Workshop on Human Computation (HCOMP)*, 2009.
25. D. Prelec. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, Oct. 2004.

26. R. Rosenthal and D. B. Rubin. Multiple contrasts and ordered bonferroni procedures. *Journal of Educational Psychology*, 76(6):1028–1034, 1984.

27. J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

28. V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. *Knowledge Discovery and Data Minding 2008 (KDD-2008)*, 2008.

29. R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, 2008.

30. L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

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

Online Labor Markets

John J. Horton

Harvard University, Cambridge, MA 02138, USA

Abstract. In recent years, a number of online labor markets have emerged that allow workers from around the world to sell their labor to an equally global pool of buyers. The creators of these markets play the role of labor market intermediary by providing institutional support and remedying informational asymmetries. In this paper, I explore market creators' choices of price structure, price level and investment in platforms. I also discuss competition among markets and the business strategies employed by market creators. The paper concludes with a discussion of the productivity and welfare effects of online labor.

1 Introduction

In the late 1990s, a number researchers began studying the effects that the Internet was having—or might yet have—on the labor market. One question examined was whether we might see the emergence of entirely online labor markets, where geographically dispersed workers and employers could make contracts for work sent "down a wire." Such markets would be an unprecedented development, as labor markets have always been geographically segmented.

Researchers were of mixed opinions: Malone predicted the emergence of such an "E-lance" market [9], while Autor was skeptical, arguing that informational asymmetries would make such markets unlikely [3]. Instead, Autor predicted the emergence of third-party intermediaries that could use their own reputation to convey "high bandwidth" information about workers—such as ability, skills, reliability and work ethic—to buyers who would be unwilling to hire workers based solely on demographic characteristics and self-reports.

In the approximately 10 years since, we have witnessed the emergence of a number of truly global online labor markets, as Malone predicted. By 2009, over 2 million worker accounts had been created across different markets, with over $700 million in gross wages paid to workers [7]. However, consistent with Autor's position, these markets have emerged not "in the wild," but within the context of highly structured platforms created by for-profit intermediaries.

The ultimate success and trajectory of these markets remains to be seen. If they become more important, they will raise policy questions, particularly about labor laws and taxation. They might spur the already large shift towards part-time employment [10] and have implications for inequality and development.

The purpose of this paper is to describe the key economic features of online labor markets. In addition to a positive examination, this paper highlights features of the markets likely to be relevant for welfare and productivity. Special

A. Saberi (Ed.): WINE 2010, LNCS 6484, pp. 515–522, 2010.

attention is given to the ability of these markets to give workers in developing countries access to buyers in rich countries.

2 Overview

Online labor markets (OLMs) fall into two broad categories: "spot" and "contest." No labor market is truly "spot" in the sense of a commodity market, but certain OLMs feature buyer/seller agreements to trade at agreed prices for certain durations of time. Examples of spot markets include oDesk, Elance, iFreelance and Guru. Workers create online profiles and buyers post jobs and wait for workers to apply and/or actively solicit applicants.

In contest markets, buyers propose contests for informational goods such as logos (e.g., 99Designs and CrowdSPRING), solutions to engineering problems (e.g., InnoCentive) and legal research (e.g., Article One Partners). Workers create their own versions of the good and the buyer selects a winner from a pool of competitors. In some markets, the buyer must agree to select (and pay) a winner before they can post a contest; in other high-stakes markets where a solution may be unlikely, the buyer is under no obligation to select a winner.

2.1 Definition

Not all people working online do so through markets: some work is unpaid (e.g., open-source software and Wikipedia) and other work products are transferred within a firm, such as through conventional off-shoring. Even within clearly identifiable markets, there is great diversity. I propose a definition of OLMs that captures the essential common features of all markets and yet distinguishes the markets from other examples of online work: a market where (1) labor is exchanged for money, (2) the product of that labor is delivered "over a wire" and (3) the allocation of labor and money is determined by a collection of buyers and sellers operating within a price system.

2.2 Nature of Labor Markets and Role for Intermediation

Labor markets are fundamentally different from other kinds of markets in at least two ways. First, there is no single "commodity" of labor with an immediately observable quality and single prevailing price—both jobs and workers are idiosyncratic. This makes it difficult for firms and workers to find a good match, and even when matches are formed, it is difficult for either party to know precisely what they are getting when they enter into contracts. Buyer/seller information asymmetries, when combined with opportunities for strategic behavior, can impede markets; if sufficiently severe, they can prevent markets from existing [2][13]. Second, labor is a service that is delivered over time, often accompanied by relationship-specific investments in human capital (e.g., learning a particular skill for a particular job), which creates a number of the incentive issues that make it hard for parties to fully cooperate [14].

In traditional labor markets, third-party intermediaries such as temp agencies, unions and testing services profit from supplying information [4]. The creators

of online labor markets do the same thing, though their scope is wider and more comprehensive. They also provide infrastructure like payment and record-keeping systems, communications infrastructures and search technology—functions typically provided by a government or by parties themselves.

2.3 What the Market Creators Provide

In order to increase the information on the demand side, OLMs often offer worker skills tests, manage reputation systems and provide worker data from prior within-OLM employment, such as hours worked and wages. Making buyer feedback public not only prevents adverse selection, but also serves to reduce moral hazard, as workers make decisions about effort "in the shadow" of the evaluations that they will likely receive. To increase supply-side information, OLM creators verify buyers' abilities to pay and reports on their past behavior in the market. For example, oDesk guarantees that workers will be paid for hourly work, putting the impetus on the buyer to interrupt an unprofitable relationship.

The influence of the market creator is so pervasive that their role in the market is closer to that of a government: they determine the space of permissible actions within market, such as what contractual forms are allowed and who is allocated decision rights.[1] Presumably they design their "institutions" to maximize expected profits. For example, they design rules to reduce the probability of disputes (subject to the constraint imposed by reducing flexibility). If disputes do arise, the market creators are likely to be able to settle them quickly using clear rules or unambiguous assignments of decision rights, such as making buyers the arbiters of contract compliance.[2]

3 Price and Price Structure

Market creators have at least three ways to earn revenue: they can charge membership fees, levy ad valorem charges on payments and charge buyers and sellers for using the market (e.g., for listing a job, taking a skills test or applying for a job).[3] These different structures are not mutually exclusive and many market creators use a hybrid structure.

A market creator has to attract both buyers and sellers to a market and facilitate valuable interactions. There is a growing literature on "two-sided" markets [12] that tries to understand price structure in the presence of membership externalities. This research focuses on scenarios where the identity of the party that

[1] Their software even serves a weights-and-measures function traditionally performed by governments by keeping universal time for logging worker hours.

[2] Although there are obvious drawbacks to such an assignment of rights, it radically reduces the space for disputes. This is in fact the precise rule used by Amazon Mechanical Turk.

[3] Typical usage fees appear to be modest and may serve as a kind of Pigouvian "tax," since some of the activities seem to be over-supplied in the markets. The costs of applying for jobs are so low that there is a good deal of application "spam."

pays the fees (or receives subsidies) matters. In online labor markets, buyers and sellers independently arrive at prices after negotiation, strongly suggesting that the Coase theorem applies, which permits a conventional one-sided analysis.[4]

Suppose that potential buyer/seller pairs would get value v from completing a project and would pay a cost c, not including any fees, if the work were intermediated. The outside option is 0. The market creator's marginal intermediation costs are assumed to be zero. If an ad valorem charge γ is leveled, the project goes forward if $v-(1+\gamma)c>0$, whereas if a lump sump fee τ is leveled, $v-c-\tau>0$. With the lump sum fee, the buyer/seller pair makes use of the market so long as the fee is less than the surplus: $\tau<v-c$. With the ad valorem charge, the pair makes use of the market so long as $\gamma<1-\frac{c}{v}$. In the lump sum case, absolute surplus matters, whereas in the ad valorem case, project efficiency matters.

Depending on the distributions of c and v, either price structure might be optimal or some hybrid might be best, but the ad valorem charge appears to have several practical advantages. First, it short-circuits the chicken-and-egg dynamics of any platform with a two-sided nature [5]. No OLM sprang forth fully formed with contingents of buyers and sellers. To be useful, the markets needed members; to attract members, they needed to be useful. An ad valorem charge avoids this problem. Second, setting an optimal lump sum charge requires knowledge of project surplus, and surplus could change dramatically as different kinds of work become more or less popular, or as firms shift more work onto the market. A firm can change membership fees, but this introduces menu costs. Finally, groups of buyers can bundle their projects under a single account and amortize their membership costs over many transactions, but this strategy offers no benefits when using usage fees. While membership fees can be important and are used in some markets, the rest of this analysis focuses on the ad valorem price structure.

3.1 Setting the Optimal ad Valorem Price Level

Perhaps because of the advantages enumerated above, ad valorem charges seem to be nearly universally applied, even in the contest markets. Assume that the buyers are purchasing efficiency units of labor from homogeneous workers and that there is a single market clearing price. The market clearing price is p and the quantity of units bought and sold is Q. Figure 1 depicts the problem in terms of intersecting supply and demand curves determining the market clearing price and quantity for a given γ. The market creator's revenue is indicated by the box with height $p\gamma$ (the side runs from p to $p(1+\gamma)$) and width Q_0. As γ grows larger, the quantity is lowered, but the height of the rectangle increases. The nested revenue box shows that as supply and demand become more elastic (S' and D'), the same size γ, even if it leads to the same market clearing price, would lead to a decrease in the quantity (and hence revenue), which is now at Q_0'.

[4] For example, if buyers have to pay a membership fee, there will be fewer buyers. This lowers the demand and therefore the price of labor, transferring some of the cost of the membership fee to sellers.

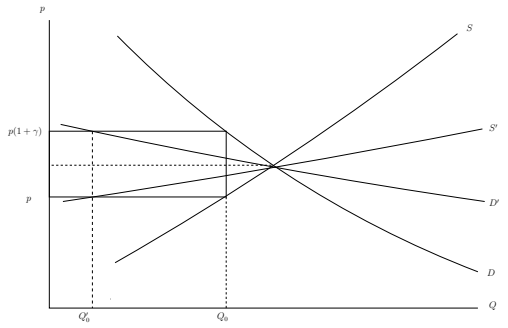


Fig. 1. Market creator profits and market clearing price

The market creator's profit maximization problem is:

$$\max_{\gamma} p(\gamma)Q(\gamma)\gamma \tag{1}$$

where γ is the ad valorem charge and $p(\gamma)$ and $Q(\gamma)$ are the resultant prices and quantities in the market. The profit-maximizing first-order condition is $(p'Q + Q'p) + pQ = 0$, which implies that profits are a maximum when $\epsilon_\gamma^p + \epsilon_\gamma^Q = -1$. The market creator increases the ad valorem charge until a small change in γ, say $x\%$ is offset by a combined $x\%$ decrease in some percentage combination of market price and quantity. Of course, the quantities ϵ_γ^Q and ϵ_γ^p are not known and are not parameters that usually receive attention in economics. However, if we make some assumptions about the functional form of the supply and demand curves, we can solve for γ^* as a function of the relevant elasticities. Assume that both the supply and demand curves have constant elasticity of substitution, $s(p) = Q_s p^\alpha$ and $d(p) = Q_d p^\beta$. In the absence of a market creator, assuming the market could function, the efficient price for labor would be $p_e = e^{\frac{\log Q_d - \log Q_s}{\alpha - \beta}}$. Under intermediation, for the market to clear, $Q_s p_I^\alpha = Q_d(p_I(1+\gamma))^\beta$. We can solve for the intermediation market clearing price, p_I, and write it in terms of the efficient market price: $p_I = p_e(\gamma+1)^{\frac{\beta}{\alpha-\beta}}$. Because $\gamma > 0$, $\beta < 0$ (downward sloping demand curve) and $\alpha > 0$ (upward sloping supply curve), in order for the market to still clear with the ad valorem charge, the price received by sellers must be lower than in the efficient market case.[5] The market creator's profits are:

$$\pi = \left[\underbrace{(Q_s p_e^\alpha) p_e}_{\text{efficient wage bill}}\right] \times \left[\gamma\,(\gamma+1)^{\frac{\beta(\alpha+1)}{\alpha-\beta}}\right] \tag{2}$$

[5] Note, however that this "distortion" from the efficient market price is not necessarily an inefficiency, as it is the actions of the market creator that make the market possible.

Solving for the optimal charge, we have:

$$\gamma^* = \frac{\beta - \alpha}{\alpha(\beta + 1)} \tag{3}$$

If we assume that the supply and demand elasticities have the same magnitude, i.e., $\alpha = |\beta|$, then in order to give the highest observed ad valorem charge of 25% employed by BitWine[6], $\alpha = |\beta| = 9$; in order to give the more standard $\approx 10\%$ used by oDesk, Amazon Mechanical Turk and others, $\alpha = |\beta| = 21$. These are remarkably high elasticities. It is not clear whether constant elasticity of substitution is a reasonable assumption, but if it is, and assuming that the market creators know their business and are not radically undercharging, it seems likely that implied elasticities are large for the simple reason that workers and buyers have ready and close substitutes for their intermediated transactions: they can make use of other online labor markets or traditional labor markets, or they can take their chances and disintermediate.

4 Competition and Specialization

It is well beyond the scope of this paper to try to model the market of intermediation markets, never mind make predictions about the likely market structure, product types, prices, etc. However, it is possible to discuss some of the key economic factors and sketch out areas for future research. The factors likely to affect ultimate market structure include whether there are economies or diseconomies of scale in providing intermediation services, barriers to entry and the potential for product differentiation.

5 Market Creator Strategy

Even after picking a price structure and level, the market creator can still increase revenues by increasing the size of the wage bill. This can be done by increasing the extent of the market, such as by recruiting more buyers and sellers, increasing worker productivity or preventing buyers and sellers from working outside the market.

5.1 Recruitment that Affects Supply and Demand

Let the market creator's initial revenues be $r_0 = p_0 Q_0$. The market creator is considering changing supply and demand in the market via recruitment of more buyers and sellers, such as through advertising. After the change, the new revenue will be $r_1 = p_1 Q_1$. Define Δx as a percentage change in x. We can write $r_1 = p_0 Q_0 (1 + \Delta p)(1 + \Delta Q)$. It is worth making the change from r_0 to r_1 if

$$\Delta Q + \Delta p + \Delta p \Delta Q > 0 \tag{4}$$

[6] BitWine is a network of freelance advisers who charge clients per-minute rates for consultations. Advisers are self-styled experts in fields such as nutrition, travel, coaching, technology and psychic prediction.

We can see that increasing within-market demand unambiguously raises profits because as ΔQ increases, so does Δp, as positive demand shocks raise both price and quantity. For supply increases, price and quantity will move in opposite directions. For small changes in supply or demand, two elasticity formulas must hold: $\Delta Q = \Delta S + \epsilon^S \Delta p$ and $\Delta Q = \Delta D + \epsilon^D \Delta p$. Because we are considering only a supply increase, $\Delta Q = \epsilon^D \Delta p$, which allows us to re-write the profit-maximizing condition as $\epsilon^D \Delta p + \Delta p + \Delta p \Delta Q > 0$. Dividing through by Δp (which is negative) and reversing the sign, we have: $\epsilon^D + \Delta Q < -1$, and since $(\epsilon^D - \epsilon^S)\Delta p = \Delta S$, the market creator finds it revenue-maximizing to increase supply so long as:

$$\epsilon_D \left(1 + \frac{\Delta S}{\epsilon^D - \epsilon^S}\right) < -1 \tag{5}$$

If supply and demand are highly elastic, $|\epsilon^D - \epsilon^S|$ is large, meaning that small positive changes in supply are likely to increase revenue.

6 Productivity and Welfare Implications

Online work offers the cost-saving benefits of telecommuting, such as reduced congestion and increased flexibility, as well as some advantages unique to the way such markets appear to be structured. First, global labor markets permit greater specialization in human capital. Second, the rapid mixture of workers across and between firms might speed up innovation spillovers, creating a kind of pseudo geographic co-location, which has been shown to increase productivity in other contexts [8]. Third, OLMs allow firms to buy small amounts of labor as needed, lowering the barriers to entrepreneurship.

OLMs also permit a kind of virtual migration that offers many of the benefits of physical migration. Assuming that increased virtual labor mobility will generate effects similar to those of increased real labor mobility, the potential gains to welfare are enormous [6]. Further, these markets create incentives for people otherwise disconnected from the global labor market to invest in their human capital [11].

Given the central role that a country's institutions play in its economic development [1], it is remarkable how little these markets demand from the institutions of the worker's home country. Prospective workers need only to be able to get online and have some way of receiving remittances. Workers do not need functioning courts, developed finance sectors, work visas, information about commodity prices, local reputations or race, class or social backgrounds required for employment in local labor markets.

Acknowledgments

Thanks to the NSF-IGERT Multidisciplinary Program in Inequality & Social Policy. Thanks to Robin Yerkes Horton, Richard Zeckhauser, Olga Rostopshova and Dana Chandler for helpful comments and suggestions.

References

1. Acemoglu, D., Johnson, S., Robinson, J.A.: Institutions as the fundamental cause of long-run growth. In: Handbook of Economic Growth,
2. Akerlof, G.: The market for "lemons": Quality uncertainty and the market mechanism. In: The Quarterly Journal of Economics, pp. 488–500 (1970)
3. Autor, D.: Wiring the labor market. Journal of Economics Perspectives 15 (2001)
4. Autor, D.H.: The economics of labor market intermediation: An analytic framework. Working Paper 14348, National Bureau of Economic Research (September 2008)
5. Caillaud, B., Jullien, B.: Chicken & egg: Competition among intermediation service providers. RAND Journal of Economics, 309–328 (2003)
6. Clemens, M., Montenegro, C., Pritchett, L., Paraguay, D., Floor, T.: The place premium: Wage differences for identical workers across the US border. World Bank Policy Research Working Paper No. 4671 (2008)
7. Frei, B.: Paid crowdsourcing: Current state & progress toward mainstream business use. Produced by Smartsheet.com (2009)
8. Greenstone, M., Hornbeck, R., Moretti, E.: Identifying agglomeration spillovers: Evidence from million dollar plants. Journal of Political Economy 118
9. Malone, T.W., Laubacher, R.J.: The dawn of the e-lance economy. Harvard Business Review 76(5), 144–152 (1998)
10. OECD. OECD Employment Outlook 2010 (2010)
11. Oster, E., Millett, B.: Do call centers promote school enrollment? Evidence from India. NBER Working Paper (2010)
12. Rochet, J., Tirole, J.: Two-sided markets: An overview. In: Institut dEconomie Industrielle Working Paper (2004)
13. Rothschild, M., Stiglitz, J.: Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. The Quarterly Journal of Economics 90(4), 629–649 (1976)
14. Williamson, O.E.: Transaction-cost economics: The governance of contractual relations. Journal of Law and Economics 22(2), 233–261 (1979)

Algorithmic Wage Negotiations: Applications to Paid Crowdsourcing

John J. Horton
Harvard University
383 Pforzheimer Mail Center
56 Linnaean Street
Cambridge, MA 02138
john.joseph.horton@gmail.com

Richard J. Zeckhauser
Harvard Kennedy School
79 JFK Street
Mailbox 41
Cambridge, MA 02138
richard_zeckhauser@harvard.edu

ABSTRACT

This paper highlights the problem that buyers and sellers encounter in arriving at prices in a distributed labor market. We argue that automated negotiation might partly remedy this problem and introduce a program, "hagglebot," that we used to negotiate payment rates for an image-labeling task with workers at Amazon's Mechanical Turk. We report the results of an initial pilot experiment that used a simple bargaining game and randomly assigned subjects to receive either a low (1 cent) or high (5 cents) offer to perform a follow-on task after they completed an initial fixed-price task. Subjects were generally reluctant to make counter-offers or end negotiations. As a result, subjects who received the 1-cent offer ended up with far lower average wages than subjects in the 5-cents group.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics; J.m [**Computer Applications**]: Miscellaneous

General Terms

Human Factors, Economics, Experimentation

Keywords

Crowdsourcing, Amazon's Mechanical Turk, Human Computation

1. INTRODUCTION

Would-be buyers and sellers of labor have to agree about wages and the quantity of work to be performed. This general problem is common to all labor markets and is often resolved through negotiation. However, negotiation is not practicable in all situations. In particular, when stakes are low, negotiation is unattractive because it is time-consuming; a prolonged negotiation can quickly dissipate any surplus that might be gained through trade. In a thick market (with many buyers and sellers for a good), posted prices and auction processes are often preferable to negotiation, and vice versa in a thin market (with few buyers and sellers). The relationship between stakes and market thickness and parties' willingness to bargain appears in many contexts: people routinely bargain over houses, cars and salaries—they (at least in the U.S.) rarely bargain over groceries or haircuts.

In distributed labor markets, the work performed by any single worker is often vanishingly small. For this reason, buyers' take-it-or-leave-it prices tend to predominate, and negotiations are rare. However, as we discuss below, posted prices have many drawbacks when used in labor markets. Negotiation might yield better net results, if it could be done more cheaply. One potentially technical way to lower the costs of negotiation would be to have it performed by a machine. An automated bot that can bargain on behalf of buyers and/or sellers has essentially no time-based opportunity cost and could help parties reach mutually beneficial agreements more efficiently.

In this note, we describe a new software tool called hagglebot that we designed to serve as an automated negotiating agent for buyers of labor. Our research using this tool is at a very early stage. This note is limited to describing hagglebot and our initial pilot study conducted in Amazon's Mechanical Turk marketplace.

1.1 Prior Work

Negotiation has received a great deal of research attention, both positive and normative [7]. Scholars and practitioners alike have noted the gap between how people actually bargain and how they should bargain. Like most scenarios in which humans have to deal with uncertainty, negotiators make a variety of systematic mistakes: they get anchored by irrelevant numbers, overweight small probabilities, are risk-seeking in the domain of losses [5], place unwarranted emphasis on the status quo [8] and so on. Even setting aside behavioral biases, negotiators make mistakes for the simple reason that negotiations are complex and cognitively demanding.[1]

One way to improve both prescriptive and descriptive knowledge is to collect more data on how people actually negotiate in real settings and, critically, to experimentally manipulate factors to determine causality. Unfortunately, creating realistic negotiation scenarios in the laboratory is challenging. However, researchers in a number of disciplines—with computer science leading the way—have begun running experiments online using online labor markets, where far greater realism is possible. Some examples in economics

[1]For this reason, there have been attempts to create decision-support tools for negotiation. Tests of some of these tools have shown that they can significantly improve outcomes [2].

CrowdConf 2010, October 4, 2010, San Francisco, CA.

include [6], [1] and [3]. Horton, Rand and Zeckhauser ([4]) argue that online experiments can offer a high degree of both internal and external validity.

2. REACHING PRICES FOR LABOR

In any potential economic exchange, both the buyer and the seller have reservation prices: for the buyer, it is the maximum amount they are willing to pay; for the seller, it is the minimum amount they are willing to accept. If the buyer's reservation price is greater than the seller's, a mutually beneficial transaction is possible.

Posit that a beneficial exchange exists. Within the range of feasible prices, buyers and sellers are locked in a zero-sum game, with any increase in price benefiting sellers and hurting buyers. As the two players jockey for a larger slice of the pie, they may make strategic demands that prevent a deal, in which case both lose.

For commodities sold in markets with many buyers and sellers, strategic considerations are moot: prices are determined by market-level supply and demand. Would-be buyers and sellers can look to the market to learn prices. In contrast, in thin markets, market prices may be unobservable and uninformative.

Labor markets can be very thin in this sense, as transactions are usually for a specific job and a specific worker. For example, there is no active market in "Joseph Jones will mow my lawn on July 15th for 1 hour," even though we can get a general sense of the price of semi-skilled manual labor. The absence of a market to reveal prices forces parties to bargain or simply accept the limitations inherent in using posted prices in thin markets.

Prices offered on take-it-or-leave-it terms are the overwhelming norm in distributed labor markets. From a buyer's perspective, setting those prices is difficult. Without knowing the distribution of workers' reservation wages for a particular task, the buyer does not know how many workers will accept a given offer; if he wishes to have n people complete the task within m days, he will be at somewhat of a loss. The problem becomes even more complex if the buyer allows workers to complete multiple tasks, as both fatigue and experience become relevant. The power to negotiate—and learn from negotiation—can help buyers avoid the pitfalls of offering too much (getting a bad deal) or too little (not getting the desired amount of work done). Through price discrimination, negotiation also has the potential to get any fixed amount of work done at an overall cheaper price.

3. HAGGLEBOT

The pioneer hagglebot is designed to negotiate over price with a worker at Amazon's Mechanical Turk (MTurk). Hagglebot can make offers, solicit counter-offers and make decisions about whether to accept counter-offers or end negotiations. It does all of these things through a natural language chat interface, with worker inputs structured in such a way that only counter-offers are entered in free-form text. While any potential bargaining strategy could be pursued by hagglebot, for experimental purposes, we have selected a simplified form of bargaining.

In our experiment, workers first completed an image-labeling task for a fixed price. After completing this initial task, workers negotiated with hagglebot over the price of labeling an additional image. (This avoided the challenge of getting workers to bid on tasks that they had not previously encountered, and hence could not estimate the difficulty in performing.)

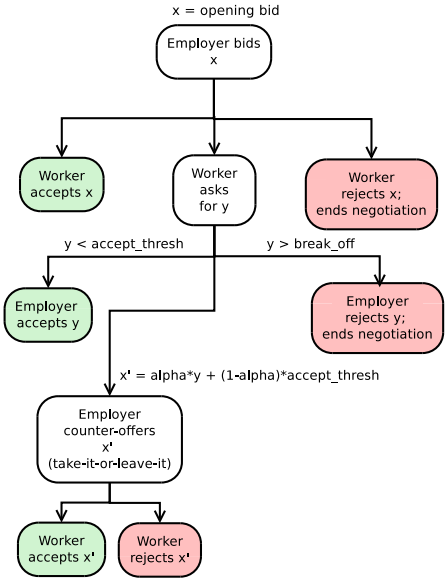


Figure 1: Restricted bargaining space

3.1 Restricted bargaining space

Unlike auctions or posted prices, negotiations can proceed in any number of ways. There are no formal rules; buyers and sellers are free to make offers, counter-offers and end negotiations. These steps can proceed in any order, and can by conditioned upon what has already transpired. The enormous "bargaining space" implied by this flexibility requires us to narrow the scope of allowed actions in order to make empirical progress, yet the narrowing must preserve realism. For the initial pilot studies, we programmed hagglebot to negotiate in a particular structure, as illustrated in Figure 1. The bargaining had at most three phases: an initial offer from hagglebot, a counter-offer from a worker and a take-it-or-leave-it counter to the worker's counter-offer from hagglebot.

3.1.1 Parameters

The buyer makes an initial offer of `opening_bid`. A worker can accept the offer, end negotiations or make a counter-offer. If `counter < accept_thresh`, work commences at the price `counter`. If `counter > break_off`, negotiations end. For values between `accept_thresh` and `break_off`, the hagglebot counter-offers

```
alpha*accept_thresh + (1-alpha)*counter
```

If the worker accepts, work commences at this new price. Otherwise, negotiations end and the new work is not undertaken.

3.2 Interface

Figure 2 illustrates the interface where workers negotiate with hagglebot. The paradigm of a chat window was

used and offers were expressed in natural language. Responses from workers were necessarily constrained to their three choices.

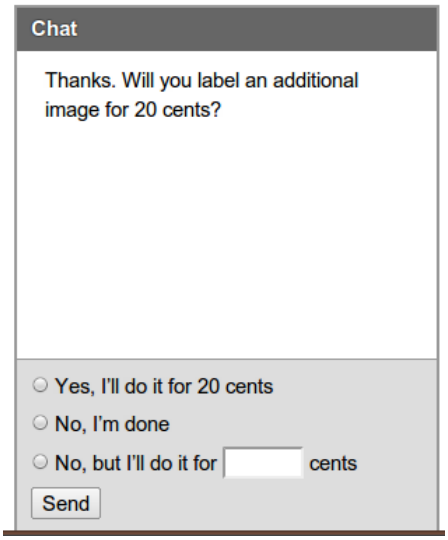


Figure 2: Chat interface

3.3 Technical Implementation

Hagglebot is written in Python and runs on the Google App Engine (GAE). Task postings to MTurk and worker payments are done using the Boto toolkit.[2] Experiment parameters are loaded into hagglebot via a structured text document containing details on payment rates, titles, group parameters, etc. Hagglebot automatically launches an experiment and collects results. The actual haggling algorithm, "hagglorithm," was written as a stand-alone script that abstracts from the details like the interface or the storage of data. This enables future researchers to easily test other bargaining strategies.

4. EXPERIMENT

For the pilot experiment, our goal was to test the software and determine whether workers are as willing to negotiate with hagglebot as they are with a human. As a paid crowdsourcing task, we used image labeling (workers were asked to give descriptive labels for an image).

1. All workers perform an initial image-labeling task for a fixed payment.

2. Workers are offered the chance to perform another image-labeling task. Negotiation is performed by hagglebot, in accordance with experimentally-assigned parameters.

3. If negotiation succeeds, work commences at the agreed price. Otherwise, negotiation ends and the second task remains unperformed.

[2]http://code.google.com/p/boto/

4.1 Experimental Design

The hagglebot parameters used for the experiment are given in Table 1. There were two experimental groups, HIGH and LOW. The two groups were identical except for the the opening bid: in HIGH, the bid was 5 cents; in LOW it was 1 cent. The experiment was run on September 6th, 2010. One hundred subjects were recruited.

Table 1: Hagglebot parameters

Paramters	HIGH	LOW
`opening_bid`	5	1
`accept_thresh`	10	10
`reject_thresh`	20	20
`alpha`	.5	.5

4.2 Results

Table 2 presents a cross-tabulation of subjects' responses to the initial offer by treatment group. A greater number of workers in LOW ended negotiations immediately after receiving offers and a greater number made counter-offers compared to what was observed in HIGH. However, a chi-square test gives a p-value of .29, suggesting that we are quite likely to get this pattern of results by chance even if there was no cross-group differences in probabilities. The main conclusion that can be drawn is that a larger sample is needed, as well as a smaller initial offer to obtain more variation in outcomes. We may also conclude that, in this example, we were well-served by starting with a low wage offer.

Table 2: Cross-tabulation of initial responses to the opening bid, by experimental group

	HIGH	LOW
Accepted offer	36	34
Ended negotiations	6	8
Made Counter offer	6	13

Notes: The chi-square test result for the cross-tabulation, with 2 degrees of freedom, is $p = 0.29$.

Because most workers accepted the opening bid regardless of group assignment, there are large differences in the mean wage earned by the different groups. Figure 3 plots the negotiated, second-task price versus experimental group. Points are jittered horizontally to prevent over-plotting. It shows that the mean price in HIGH was considerably higher than the mean price in LOW, with no overlap in the 95% confidence intervals. Confirming what is graphically clear, the regression line is:

$$price = \underbrace{-2.991}_{[0.606]} LOW + \underbrace{5.791}_{[0.311]} \quad (1)$$

with $N = 93$ and $R^2 = 0.2$.

5. CONCLUSION

This note introduces a planned program of research aimed at understanding how workers negotiate and how this knowledge can be used in applications. The simple pilot described

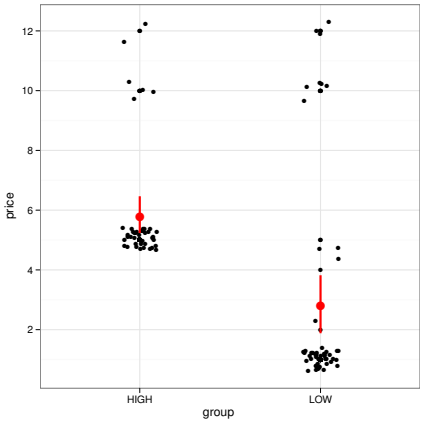


Figure 3: Agreed wage (per image) for successful negotiations (points horizontally jittered) .

already illustrates, for example, the advantages of more flexible pricing: wages were considerably lower in LOW without inducing significant differences in uptake. Potential future extensions include testing different parameter configurations, examining actual label output and conditioning worker behavior on observable characteristics. Hagglebot might then be modified to allow for bargaining strategies that adapt in real time to reflect knowledge gained from the outcome of bargaining.

6. REFERENCES

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

[2] E. Chen, R. Vahidov, and G. Kersten. Agent-supported negotiations in the e-marketplace. *International Journal of Electronic Business*, 3(1):28–49, 2005.

[3] J. Horton and L. Chilton. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, 2010.

[4] J. Horton, D. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *NBER Working Paper w15961*, 2010.

[5] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47(2):263–291, 1979.

[6] W. Mason and D. Watts. Financial incentives and the 'performance of crowds'. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, 2009.

[7] H. Raiffa. *The art and science of negotiation*. Belknap Press, 1982.

[8] W. Samuelson and R. Zeckhauser. Status quo bias in decision making. *Journal of risk and uncertainty*, 1(1):7–59, 1988.

Task Search in a Human Computation Market

Lydia B. Chilton
University of Washington
hmslydia@cs.washington.edu

John J. Horton
Harvard University
horton@fas.havard.edu

Robert C. Miller
MIT CSAIL
rcm@mit.edu

Shiri Azenkot
University of Washington
shiri@cs.washington.edu

ABSTRACT

In order to understand how a labor market for human computation functions, it is important to know how workers search for tasks. This paper uses two complementary methods to gain insight into how workers search for tasks on Mechanical Turk. First, we perform a high frequency scrape of 36 pages of search results and analyze it by looking at the rate of disappearance of tasks across key ways Mechanical Turk allows workers to sort tasks. Second, we present the results of a survey in which we paid workers for self-reported information about how they search for tasks. Our main findings are that on a large scale, workers sort by which tasks are most recently posted and which have the largest number of tasks available. Furthermore, we find that workers look mostly at the first page of the most recently posted tasks and the first two pages of the tasks with the most available instances but in both categories the position on the result page is unimportant to workers. We observe that at least some employers try to manipulate the position of their task in the search results to exploit the tendency to search for recently posted tasks. On an individual level, we observed workers searching by almost all the possible categories and looking more than 10 pages deep. For a task we posted to Mechanical Turk, we confirmed that a favorable position in the search results do matter: our task with favorable positioning was completed 30 times faster and for less money than when its position was unfavorable.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; J.4 [**Social and Behavioral Sciences**]: Economics

General Terms

Economics, Experimentation, Human Factors, Measurement

KDD-HCOMP'10, July 25th, 2010, Washington,DC, USA.

Keywords

Search, Amazon Mechanical Turk, Human Computation, Crowdsourcing, Experimentation

1. INTRODUCTION

In every labor market, information plays a critical role in determining efficiency: buyers and sellers cannot make good choices unless they know their options. According to a rational economic model of labor supply, workers use information about the availability and the nature of tasks to make decisions about which tasks to accept. If workers lack full information about the tasks, they are likely to make suboptimal decisions, such as accepting inferior offers or exiting the market when they would have stayed, had they known about some other task. In large markets with many buyers and sellers, lack of knowledge about all available options is a key source of friction. This lack of knowledge stems from the fact that searching is neither perfect nor costless.

The "search problem" is particularly challenging in labor markets because both jobs and workers are unique, which means that there is no single prevailing price for a unit of labor, nevermind a commodified unit of labor. In many labor markets, organizations and structures exist to help improve the quality and quantity of information and reduce these "search frictions." We have already seen conventional labor markets augmented with informational technology, such as job listing web sites like Monster.com [2], and yet a great deal of of information is privately held or distributed through social networks and word of mouth.

In online labor markets, there is relatively little informal sharing of information as in traditional markets. Given that online labor markets face no inherent geographic constraints, these markets can be very large, further exacerbating search frictions. These frictions can and are eased by making the data searchable through a variety of search features. If agents can sort and filter tasks based on characteristics that determine the desirability of a task, they are able to make better choices, thereby improving market efficiency. Of course, this hypothetical gain in efficiency depends upon the search technology provided and the responses of buyers to that technology.

Amazon's Mechanical Turk (MTurk) is an online labor market for human computation. It offers several search features that allow workers to find Human Intelligence Tasks (HITs). HITs can be sorted by fields such as most available, highest reward, and title. HITs are also searchable by keyword, minimum reward, whether they require qualifica-

tions (and combinations thereof). In this paper, we seek to provide insight into how workers search for tasks.

We present the results of two methods for investigating search behavior on MTurk. First, we scrape pages of available HITs at a very high rate and determine the rate at which a type of HIT is being taken by workers. We use a statistical model to compare the rate of HIT disappearance across key categories of HITs that MTurk allows workers to search by. Our premise for observing search behavior is that search methods which return HITs with higher rates of disappearance are the search methods which workers use more. This method relies only on publicly available data—the list of HITs available on www.mturk.com. Second, we issued a survey on MTurk asking over 200 workers how they searched for tasks. We posted the survey with carefully chosen parameters so as to position it so that it can be found more easily by some search methods and not others. This way, we can target search behaviors that are not picked up by the analysis of the scraped data.

In this paper, we first motivate studying search behavior in an online human computation labor market by reviewing related work. We next describe MTurk's search features and our method of data collection based on these search features. We then present a statistical analysis of our high-frequency scraped data followed by the results of a survey of MTurk workers.

2. RELATED WORK

In the field of human computation, it is important to ask "Why are people contributing to this effort?" What are the motivations behind labeling an image [17], uploading videos to YouTube [11], or identifying the genre of a song [13]? On MTurk, the main motivation seems to be money [12, 10]. However, financial motivations do not imply that workers are only motivated to find HITs offering the highest rewards: workers may choose from among thousands of HITs that differ in required qualifications, level of difficulty, amount of reward, and amount of risk. With all of these factors at play, workers face significant decision problems. Our task was to model this decision problem.

Because MTurk is a labor market, labor economics—which models workers as rational agents trying to maximize net benefits over time—may provide a suitable framework. Research on the labor economics of MTurk has shown that workers do respond positively to prices but that price is not indicative of work quality [14]. Previous work [9] shows that not all workers follow a traditional rational model of labor supply when choosing their level of output. Instead of maximizing their wage rate, many workers create target earnings—they focus on achieving a target such as $.25 or $1.00 and ignore the wage rate. Although price is important to task choice, it is clearly not the only factor; research on search behavior is necessary in order to more fully understand how workers interact with the labor market.

In other domains where users must cope with a large information space, progress and innovation in search have had enormous positive impacts. Developments such as PageRank [4] have made it possible to search the long tail of large, linked document collections by keyword. Innovations such as faceted browsing [8] and collaborative filtering [7] have made it faster and easier to search within more structured domains such as shopping catalogues. Web queries are sensitive to changes in content over time, and searchers have an increasing supply of tools to help them find updates on webpages they revisit [1]. In labor and human computation markets, search tools are still evolving and it remains to be determined what features will prove most useful in overcoming search frictions.

Web search analysis has shown that the first result on the first page of search results has by far the highest clickthrough rate [16]. In a market model where parties have full information, there would be no need for search technologies—no purely "positional effects"—all parties would instantly and costlessly find their best options. In reality, search technologies are likely to have strong effects on the buyer-seller matches that are made. This point is evident by the fact that companies are willing to spend billions of dollars to have advertisements placed in prime areas of the screen [5].

While there are similarities between many kinds of online search activities, we must be careful in generalizing. Search behavior in one domain does not necessarily carry over to other domains. In this paper we focus solely on the domain of workers searching for HITs on MTurk, using web search behavior as a guide to what features may be important.

3. MTURK SEARCH FEATURES

At any given time, MTurk will have on the order of $100,000$ HITs available. HITs are arranged in HIT groups. All HITs posted by the same person with the same featues (title, reward, description, etc) are listed together in a HIT group to avoid repetition in the list of available HITs. HIT groups are presented much like traditional web search engine results, with 10 HIT groups listed on each page of search results. By default, the list is sorted by "HITs Available (most first)." The default view gives seven pieces of information:

1. Title (e.g., "Choose the best category for this product")
2. Requester (e.g., "Dolores Labs")
3. HIT Expiration Date (e.g., "Jan 23, 2011 (38 weeks)")
4. Time Allotted (e.g., "60 minutes")
5. Reward (e.g., "$0.02")
6. HITs Available (e.g., 17110)
7. Required qualifications (if any)

By clicking on a HIT title, the display expands to show three additional fields:

1. Description (e.g., "Assign a product category to this product")
2. Keywords (e.g., categorize, product, kids)
3. Qualifications. (e.g., "Location is US")

The MTurk interface also offers several search features prominently positioned at the top of the results page, including a keyword search and a minimum reward search. Additionally, HITs can be sorted in either ascending or descending order by six categories:

1. HIT Creation Date (***newest*** or ***oldest***)
2. HITs Available (***most*** or ***fewest***) (i.e., how many subtasks may be performed)

3. Reward Amount (***highest*** or ***lowest***)

4. Expiration Date (***soonest*** or ***latest***)

5. Title (***a-z*** or ***z-a***)

6. Time Allotted (***shortest*** or ***longest***)

Some sort results change more quickly than others. Sorting by most recent creation date will produce a very dynamic set of HIT groups. However, sorting by most HITs Available will produce a fairly static list because it takes a long time for a HIT group that contains, say, 17,000 HITs to fall off the list.

On each worker's account summary page, MTurk suggests 10 particular HITs, usually with high rewards but no other apparent common characteristic. Other than this, we are unaware of any services to intelligently recommend HITs to workers. We presume that nearly all searching is done using the MTurk search interface. A Firefox extension called Turkopticon [15] helps workers avoid tasks posted by requesters that were reported by other users as being unfair in their view. This could affect workers' choices of which tasks to accept, but probably has little effect on the category workers choose to sort by and does not reorder or remove HITs from the results pages.

Requesters post tasks on MTurk and pay workers for acceptable work. There are two basic types of HITs that encompass most postings: (1) tasks such as image labeling, where workers are invited to perform as many various tasks as are available to be completed, and (2) tasks such as surveys, which require many workers but allow each worker to do the HIT only once. Type (2) HIT groups appear to workers as single HIT because there is only one HIT available to each worker. If a worker does not accept the HIT, it will remain on his list of available HITs until the total number of performances desired by the requester are completed. It often takes a substantial amount of time to achieve the throughput of workers necessary to complete type (2) HITs where the task only appears as a single available HIT to each worker. Because it is a small wonder they those HITs get done at all, we call them "1-HIT Wonders."

For HIT groups that allow individual workers to perform a single task multiple times (type (1)), the number of "HITs Available" is constantly updated. The number can increase in only two cases: first, when a HIT is returned by a worker unperformed and, second, in the rare event that more tasks are added to the HIT group before it is finished. Fortunately, this second case is easy to detect.

In this paper, we use HIT disappearance rates to perform a statistical analysis of search behavior for type (1) HIT groups containing multiple tasks available to all workers. We complement our analysis with the results of a survey where workers were asked about how they search for tasks such as 1-HIT Wonders that the scraping does not address.

4. METHOD A: INFERENCES FROM OBSERVED DATA

We want to determine whether the sort category, the page on which a HIT appears (page placement), and the position it occupies on that page (page position) affect the disappearance rate of a HIT. Our premise is that if any of these factors affect the disappearance rate of a HIT, then workers are using that search feature to find HITs. This approach makes use of what economists call a "natural experiment".

To understand it, consider the following thought experiment: Imagine that we could randomly move HITs to different pages and page positions within the various search categories without changing the attributes of the HITs. With this ability, we could determine the causal effects of HIT page placement and page position by observing how quickly a HIT is completed. Of course, we do not actually possess the ability to manipulate search results in this way for HITs we do not post, but we are able to observe the functioning of the market and potentially make a similar inference based on the scraped data.[1]

The problem with this approach, however, is that HIT characteristics that are likely to affect the popularity of a HIT should also affect its page placement and page position within the search results of the various sorting categories. In fact, the relationship between attributes and search results page placement and page position is the whole reason MTurk offers search features—if search results page placement and page position were unrelated to HIT characteristics that workers cared about, then any search category based on those characteristics would offer no value.

To deal with the problem of correlation between attributes and page placement and page position, we needed to make several assumptions and tailor our empirical approach to the nature of the domain. Before discussing our actual model, it will be helpful to introduce a conceptual model. We assume that the disappearance of a particular HIT during some interval of time is an unknown function of that HIT's attributes and its page placement and page position (together, its position in the list of search results):

$$\text{Disappearance of HIT } i = F(\text{position of } i, X_i)$$

where X_i are all the HIT attributes, market level variables, and time-of-day effects that might also affect disappearance rate. Our key research goal was to observe how manipulations of the position of i—while keeping X_i constant—affect our outcome measure, the disappearance of HITs (i.e., work getting done). We were tempted to think that we could include a sufficient number of HIT co-variates (e.g., reward, keywords, etc.) and control for the effects of X_i, but this is problematic because we cannot control for all factors.

Our goal of untangling causality was complicated by the fact that we cannot exogenously manipulate search results and the fact that there are HIT attributes for which we cannot control. Our approach was to acknowledge the existence of unobserved, HIT-specific idiosyncratic effects but to assume that those effects are constant over time for a particular HIT group. With this assumption, we relied upon movement in HIT position within a sort category as a "natural experiment." When we observed a HIT moving from one position to another, and then observed that the HIT disappeared more quickly in the new position, we attributed the difference in apparent popularity to the new position, and not to some underlying change in the nature of the HIT.

4.1 Data Collection

There are 12 ways in which MTurk allows workers to sort HITs—six categories, each of which can be sorted in ascend-

[1]In Method B, we will actually manipulate our HITs position in the search results (albeit by modifying the HIT attributes).

ing or descending order. We scraped the first three pages of search results for each of the 12 sorting methods. Each page of results lists 10 hits, which meant that we scraped 360 HITs in each iteration. We scraped each of the 36 pages approximately every 30 seconds—just under the throttling rate. Although we ran the scraper for four days, due to our own data processing constraints, in this analysis we used data collected in a 32-hour window beginning on Thursday, April 29, 2010, 20:37:05 GMT and ending on Saturday, May 1, 2010, 04:45:45 GMT. The data contains 997,322 observations. Because of the high frequency of our scraping, each posted HIT is observed many times in the data: although the data contains nearly 1,000,000 observations, we only observed 2,040 unique HITs, with each HIT observed, on average, a little more than 480 times.

For each HIT, we recorded the 10 pieces of information that the MTurk interface offers (seven default and three additional features—see Section 3), as well as the sort category used to find the HIT, the page placement of the HIT (1, 2 or 3), and the page position of the HIT (first through tenth).

4.1.1 Measuring HIT disappearance

Formally, let s index the sort category, let g index the groups of observations of the same HIT (which occur because each HIT is observed many times), and let i index time-ordered observations within a HIT group. Let HITs be ordered from oldest to most recently scraped, such that $i+1$ was scraped after i. Let the number of HITs available for observation i be y_{igs}. The change is simply $\Delta y_{igs} = y_{(i+1)gs} - y_{igs}$.

Unfortunately, there are several challenges in treating this outcome measure as a direct measure of uptake. First, requesters can add and delete tasks in a HIT group over time. If a requester deletes a large number of tasks, then a regression might incorrectly lead us to believe that there was something very attractive to workers about the page placement and/or page position of the HIT. Second, for 1-HIT Wonders, the HITs-available measure might not change even though the HIT is very popular. Finally, the absolute change is "censored" in that the number of HITs available cannot go below zero, which means that HIT groups with many tasks available are able to show greater changes than HIT groups with few tasks available. For these reasons, we make our outcome variable an indicator for a drop in HITs available, regardless of the magnitude of the change:

$$Y_{igs} = 1 \cdot \{\Delta y_{ijs} < 0\}$$

4.2 Econometric set-up

To implement our conceptual model, we assumed a linear regression model in which the expected outcome is a linear function of a series of variables. Obviously, many factors determine whether or not a HIT disappears from search results. Page placement, page position, and sort category certainly matter, but so do variables that are hard to observe and which we cannot easily include in a regression. For example, a comparatively high-paying, short task that looks fun will probably be accepted more quickly than one that is low-paying, long, and tedious. In creating our model, we sought to separate the desirability of a task from the page placement and page position it occupies in the search results of a particular sorting category.

4.2.1 Choosing a model

Rather than simply introduce a number of controls that will invariably miss factors that are unmeasured, we included a HIT group-specific control in the regressions called a "group random effect." Each HIT group was modeled as having its own individual level of attractiveness that stays constant over time. In this way, all factors that have a static effect on HIT attractiveness were controlled for, regardless of whether or not we are able to observe the factors that determine a HIT groups attractiveness.

With a group-specific random effect, we eliminate one of they key problems that would arise from simply using a pooled model. For example, a fun HIT that always stays on the third page of results sorted by highest reward might cause us to erroneously believe that the third page is very important, when the case is simply that one HIT group is disappearing quickly.

One necessary word on the nomenclature of these two models: when dealing with data that has a natural group structure, a model with a group specific-effect is called a "random effects" model or "multilevel" model; a model that ignores the group structure and does not include any group-specific effects is called a "pooled" model. Both types are linear regression models. While our group-specific effect model "controls" for group factors, we do not actually include a dummy variable for each group (which is called the "fixed effects" model). Rather, we start with a prior assumption that each effect is a random draw from a normal distribution (the random effects model) [6]. As the number of observations increases, this fixed effects/random effects distinction becomes less important.

4.2.2 Model

The group-specific random effects model is as follows: for an observation i of group g, we estimate

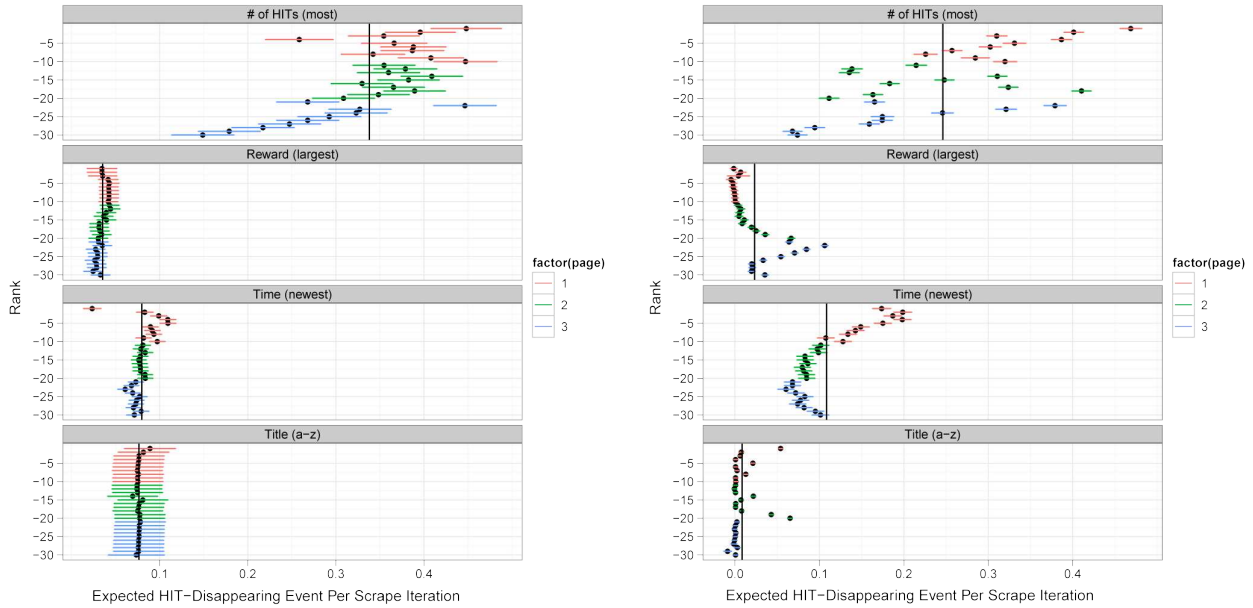
$$Y_{igs} = \sum_{r=1}^{30} \beta_s^r x_{igs}^r + \gamma_g + \tau_{H(i,g)} + \epsilon \qquad (1)$$

where $x_{ig}^r = 1$ if the HIT i is at position r (and $x_{ig}^r = 0$ otherwise) and where $\gamma_g \sim N(0, \sigma_\gamma^2)$ is the group-specific random effect and $\tau_{H(i)} \sim N(0, \sigma_\tau^2)$ is a time-of-day random effect. The pooled model ignores this grouped structure and imposes the constraint that there is no group effect and $\gamma_g = \gamma = 0$.

4.3 Results

We applied our models to four of MTurk's 12 sorting options: *newest* HITs, *most available* HITs, *highest reward*, and *title a-z*. The others—including *shortest* time allotted and *latest expiration*—tend to produce 1-HIT Wonders and don't seem like natural ways to sort HITs.

In Figure 1, the collection of coefficients, β_s^r, from Equation 1 are plotted for both the group-specific random effects model (panel (a)) and the pooled model (panel (b)), with error bands two standard errors wide. The four sorting options are listed at the top of the figure: *newest*, *most*, *highest reward*, *a-z*. The points are the coefficients (with standard error bars) for each page and position on that page, ranging from one to 30 (positions are displayed as negative numbers to preserve the spatial ordering). The coefficients $\hat{\beta}_s^r$ are interpretable as the probability that a HIT occupying position r decremented by one or more HITs during the scraping



(a) Group-specific random effects model

(b) Pooled model

Figure 1: Effects of position in sort category on HIT disappearance. Pages are illustrated with color. Horizonal lines represent one standard error on either side.

interval. Our premise was that sort categories, page placements, and page positions with higher probability of the HIT decrement are those that workers search by.

4.4 Group random effects model

Under the assumptions of the group-specific random effects model, we interpret the position coefficients as the probability that a HIT occupying a particular page position will have a disappearance event during the scraping interval. The only sort category that shows strong positional effects is *most*, with similar rates for the first and second results pages, and then a strong drop-off on the third page.

From this, we conclude that workers actively sort by the *most* available HITs and look at results on the first two pages of search results. On the third page of *most* available HITs, the average coefficient is high, but position effects seem to matter. We conjecture this to mean that the further down the third page the HIT is located, the more likely workers are to abandon this search strategy. However, more investigation is needed in order to confirm this interpretation.

The page placements and page positions generated by the *highest reward* sorting category have no apparent effect on search behavior. The overall levels of disappearance for HITs among the search results are low, which is to be expected given the relative unpopularity of high-reward HITs. Also unsurprisingly, the page placements and page positions generated by the *title a-z* sorting category have no apparent effect on search behavior.

4.5 Pooled model

The group-specific random effects results shown in panel (a) of Figure 1 are more credible than those of the pooled model in panel (b). Nevertheless, panel (b) coefficients are far more precise. This is because panel (a) coefficients are determined solely by within-group movements in search result position. For sorting categories that do not show much movement—such as *title*—the panel (a) estimates are comparatively imprecise. However, the panel (b) estimates only capture the innate attractiveness of whatever HIT was occupying a particular position. For example, at $r = 20$ in panel (b), *title*, we see a strong statistically significant effect, but this is presumably only a feature of whatever HIT was occupying that position.

This "stationary HIT" bias issue also arises in panel (b), *reward*: on the first page of search results for *highest*, there is almost zero probability of the HIT disappearing. It may seem surprising that high-reward tasks aren't being taken, but this fact can be readily explained. Most MTurk HITs have a reward between \$0.02 and \$1.00. However, there are some HITs that get posted for larger sums, on the order of \$10.00. There are very few such HITs and they tend to be unpopular, so they stay on the *highest reward* page for a long time without disappearing. We do see that on the second and third results pages, the probability of non-zero disappearance rate goes up.

4.5.1 *Searching by newest HITs?*

According to panel (a), the search results position within *newest* HITs is generally either irrelevant, or, in the case of position 1, page 1, is actually very harmful! This *is* surprising. Sorting by the *newest* HITs seems to be a very good way to find fresh and interesting tasks. Further, as our survey results (described below) show, most workers report searching for HITs based *newest*.

Interestingly, *newest* is a sort category where we would expect the inclusion of group effects to be irrelevant, since position in the *newest* sort category should be essentially mechanical—a HIT group is posted, starts off at the first position, and then moves down the list as newer HIT groups are posted. Yet, in comparing *newest* across the two panels, the effects of position are radically different: in panel (b), we see that position has a strong effect on uptake, with the predicted pattern, while in panel (a), there appears to be little of no effect. To summarize, the pooled model suggests strong effects, while the random effects model suggests no effects.

4.5.2 Discrepancy

We hypothesized that there were no positional effects in the group-specific random effects model because certain requesters actively game the system by automatically re-posting their HIT groups in order to keep them near the top of *newest*. This hypothesis reconciles two surprising findings.

First, gaming explains the absence of *newest* effects in panel (a): the benefits to being in a certain position are captured as part of the group effect. To see why, consider a requester that can always keep his HIT at position 1. No matter how valuable position 1 is in attracting workers, all of this effect will be (wrongly) attributed to that particular HIT's group-specific effect.

Second, gaming explains why position 1 appears to offer far worse performance in panel (a) (and why there is a generally increasing trend in coefficient size until position 4). A very large share of the HITs observed at position 1 in the data came from gaming requesters. A still large but relatively smaller share of the HITs observed at position 2 came from gaming requesters who were trying to get their HIT groups back into position 1, and so on. Because all the purely positional benefits get loaded onto the group effects of gaming requestors, positions more likely to be occupied by gaming requestors will appear to offer smaller positional benefits.

Gaming not only reconciles the data—it is also directly observable. We identified the HITs that had 75 or more observations in position 1 of *newest*. In Figure 2, the position $(1-30)$ is plotted over time with the HIT group ID and title posted above. It was immediately clear that some HITs are updated automatically so that they occupy a position near the top of the search results: until day 0.5, the HIT groups in the second and third positions were "competing" for the top spot; around day 0.9, they faced competition from a third HIT group (first panel) that managed to push them down the list.

Under these gaming circumstances, the group-specific random effects model is inappropriate, as the positions are not randomly assigned. For this reason, the pooled model probably offers a clearer picture of the positional effects, though the coefficient estimates are certainly biased because they incorporate the nature of HITs occupying position one, which we know is from gaming.

We conclude that the pooled regression results are the correct analysis for the *newest* HITs category. The pooled regression results show that workers are actively searching by *newest* HITs and focus on the first page of results.

5. METHOD B: WORKER SURVEY

Another way to study how workers search for HITs is to ask the workers directly, by conducting a survey on MTurk. Although this approach is more expensive, more obtrusive, and smaller-scale than scraping the MTurk web site, it has two advantages. First, surveying complements the scraping technique. Scraping is better suited to tracking postings of large numbers of type (1) HITs whose tasks can be performed multiple times by any worker, so an anonymous scraper can observe their acceptance. Surveys, by contrast, are necessarily 1-HIT Wonders, offered at most once to every worker, and MTurk offers no way for a scraper to observe how many workers have accepted the offer. Since a substantial use of MTurk involves 1-HIT Wonders for online experimentation and surveys of other kinds, we need a method for exploring search behavior on this side of the market. Second, surveys can provide more detailed and subjective explanations for worker behavior than aggregate statistics.

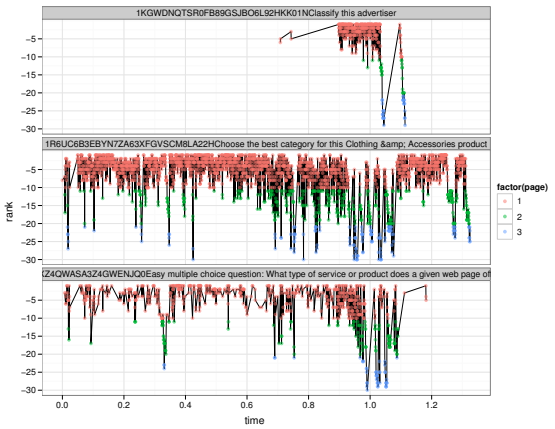


Figure 2: Position of three HITs over time sorted by *newest*. (Time measured in days)

This section presents the results of a survey of roughly 250 workers about how they sort and filter HITs. Since MTurk was used to conduct the survey, the characteristics of the HIT used to solicit survey respondents have an effect on the results. We explored this effect by posting the survey in several different ways in order to sample workers who exhibited different search behaviors. In the process, we observed how certain choices of HIT characteristics can make a HIT very hard to find, substantially reducing the response rate. Finally, the survey also collected free-form comments from workers about how they find good and bad HITs.

5.1 Procedure

The survey was designed to probe respondents' immediate search behavior—specifically, how respondents used MTurk's search interface to discover the survey HIT itself. The survey asked several questions: (1) which of the 12 sort categories they were presently using; (2) whether they were filtering by a keyword search or minimum price; and (3) what page number of the search results they found the survey HIT on. A final question asked the worker for free-form comments about how easy or hard it is to find HITs.

The survey was posted on MTurk in four ways, using different HIT characteristics for each posting to give it high and low positions in the various sorted search results. Aside

from these manipulated parameters, the four postings were identical (e.g. in title, description, preview, and actual task).

Best-case posting. In one posting, the parameters of the HIT were chosen to place it among the top search results under each of the six primary sort categories. Thus the HIT would appear on the first page, or as close as possible, when HITs were sorted by *any* attribute, though only in one direction (ascending or descending). The goal of this posting was to capture as many workers as possible from each of the six chosen sort orders that the posting optimized.

The HIT was automatically one of the *newest*, at least at first, and had the *fewest* possible HITs available because it was only offered once to each worker. For reward, we chose the *least* amount: $0.01. For the *soonest* expiration date and the *shortest* time allotted, we chose 5 hours and 5 minutes, respectively. Finally, we optimized for alphabetical title order, *a-z*, by starting the title with a space character.

The best-case posting was also labeled with the keyword "survey", which we knew from pilot studies was frequently used by turkers to find desirable survey HITs. The other three postings had the same title and description as the best-case posting, but no "survey" keyword.

Worst-case posting. To test the impact of especially poor position on a HIT's reachability, we also posted the survey in a way that was *hard to find* as possible for workers using any sorting category. This was done by choosing parameter values near the median value of existing HITs on MTurk, so that the survey HIT would appear in the middle of the pack, as far as possible from the first page in both ascending and descending order by that attribute.

For example, at the time the survey was posted, the median number of HITs available was two. Postings with two HITs covered pages 55–65 (out of 100 pages of HITs requiring no qualifications) when the list was sorted by *fewest* HITs available, and pages 35–45 when sorted by *most* HITs. As a result, a worker sorting in either direction would have to click through at least 35 pages to reach any 2-HIT posting, which is highly unlikely. We therefore posted the survey with two independent HITs.

For the remaining attributes, we chose a reward amount of $0.05 and an expiration period of one week and allotted 60 minutes in order to position the HIT at least 20 pages deep among search results sorted by reward amount, creation date, and time allotted, in both ascending and descending order. The median title of all current HITs started with "Q,", so we started the survey title with the word "Question" to make it hard to find whether sorting by *a-z* or *z-a*.

Creation date, however, cannot be chosen directly. In particular, a HIT cannot be posted with a creation date in the past. Freshly posted HITs generally appear on the first page of the *newest* sort. In order to artificially age our survey HIT before making it available to workers, we initially posted a nonfunctional HIT, which presented a blank page when workers previewed it or tried to accept it. After six hours, when the HIT had fallen more than 30 pages deep in the *newest* sort order, the blank page was replaced by the actual survey, and workers who discovered it from that point on were able to complete it.

Newest-favored and a-z-favored posting. We performed two additional postings of the survey, which were intended to favor users of one sort order while discouraging other sort orders. The *newest-favored* posting was like the worst-case posting in all parameters except creation date. It appeared functional immediately, so users of the *newest* sort order would be most likely to find it. Similarly, the *a-z-favored* posting used worst-case choices for all its parameters except its title, which started with a space character so that it would appear first in *a-z* order.

All four postings were started on a weekday in May 2010, with their creation times staggered by 2 hours to reduce conflict between them. The best-case posting expired in 5 hours, while the other three postings continued to recruit workers for 7 days.

5.2 Results

Altogether, the four postings recruited 257 unique workers to the survey: 70 by the best-case posting (in only 5 hours), 58 by the worst-case posting, 59 by newest-favored, and 70 by a-z-favored (all over 7 days). Roughly 50 workers answered more than one survey posting, detected by comparing MTurk worker IDs. Only the first answer was kept for each worker.

Figure 3 shows the rate at which each posting recruited workers to the survey over the initial 24-hour period. The best-case posting shows the highest response rate, as expected, and the worst-case posting has the lowest. The newest-favored posting shows a pronounced knee, slowing down after roughly 1 hour, which we believe is due to being pushed off the early pages of the *newest* sort order by other HITs. Because the posting's parameters make it very hard to find by other sort orders, it subsequently grows similarly to worst-case. The a-z-favored posting, by contrast, recruits workers steadily, because it remains on the first page of the a-z sort order throughout its lifetime. These results show that HIT parameters that affect sorting and filtering have a strong impact on the response rate to a HIT. All four postings were identical tasks with identical descriptions posted very close in time, and the best-case posting actually offered the lowest reward ($0.01 compared to $0.05 for the others), but its favorable sort position recruited workers roughly 30 times faster than the worst-case posting.

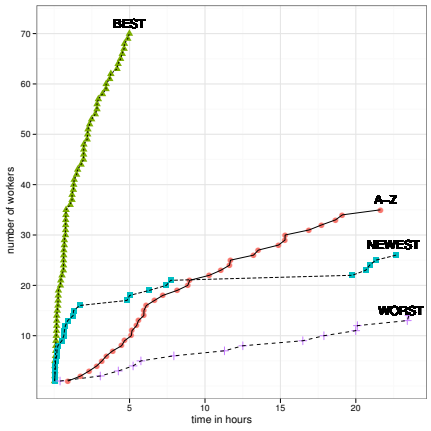


Figure 3: Number of workers responding to each survey posting within the first 24 hours of posting (time measured in days)

Figure 4 shows the responses to the survey question about sorting category, grouped by the posting that elicited the re-

sponse. The best-case posting recruited most of its workers from the *newest* sort order, reinforcing the importance of this sort order for 1-HIT wonders like a survey. The *newest* sorting category also dominated the responses to the newest-favored posting, as expected, and *a-z* appeared stongest in the a-z-favored posting. Strong appearance of *most* HITs was a surprise, however, since each posting had only 1 or 2 HITs.

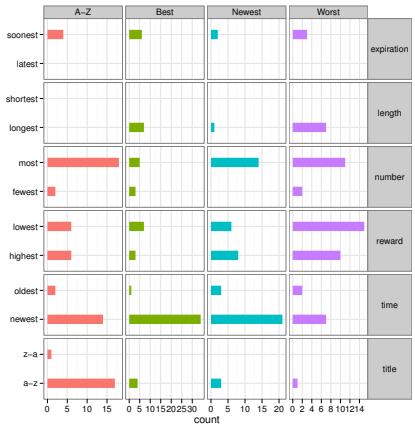


Figure 4: Number of workers who reported using each sort order to find each of the four postings.

For the survey question about reward amount filtering, roughly 65% of the survey respondents reported that they had not set any minimum amount when they found the survey HIT. Surprisingly, however, 9% reported that they had used a reward filter that was *greater* than the reward offered by the survey HIT. For example, even though the best-case posting offered only $0.01, ten respondents reported that they found the survey while searching for HITs that paid $0.05 or more. We followed up by sending email (through MTurk) to several of these workers, and found that the problem can be traced to a usability bug in MTurk: *"Once we submit a hit MTurk takes us back to all hits available, 1st page with no $ criteria."* In other words, a filter is discarded as soon as the worker performs a single task. Several other workers mentioned similar problems in their free-form comments.

Similarly, roughly 20% of the workers reported using a keyword filter when they found the survey, but half of these reported using keywords that were not actually included in the survey HIT, suggesting that the filter they thought they were using was no longer in effect. 21 workers reported using the "survey" keyword, of whom 16 were responding to the best-case posting, which actually did include that keyword.

Figure 5 shows the distribution of result page numbers on which workers reported finding the survey HIT. Roughly half found it on the first or second page of results, as might be expected. Yet a substantial fraction (25%) reported finding the survey beyond page 10. Because of the way the survey was posted, particularly the worst-case posting, workers would not have been likely to find it anywhere else. Yet it is still surprising that some workers are willing to drill dozens of pages deep in order to find tasks.

The free-form comment question generated a variety of feedback and ideas. Many respondents reported that they would like to be able to search for or filter out postings from particular requesters. Other suggestions included the ability to group HITs by type, and to offer worker-provided ratings of HITs or requesters. Several workers specifically mentioned the burden of clicking through multiple pages to find good HITs:

"Scrolling through all 8 pages to find HITs can be a little tiring."

"I find it easier to find hits on mechanical turk if I search for the newest tasks created first. If I don't find anything up until page 10 then I refresh the page and start over otherwise it becomes too hard to find tasks."

Several workers pointed out a usability bug in MTurk that makes this problem even worse:

"It is difficult to move through the pages to find a good HIT because as soon as you finish a HIT, it automatically sends you back to page 1. If you are on page 25, it takes so much time to get back to page 25 (and past there)."

"Please keep an option to jump to a page directly without opening previous pages."

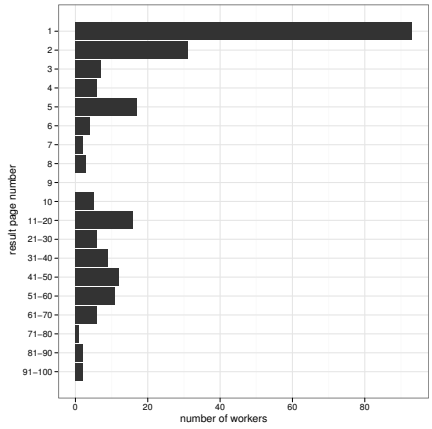


Figure 5: Histogram of result page numbers on which workers found the survey HIT. Most found it on page 1, but many were willing to drill beyond page 10.

6. CONCLUSION

In this paper, we studied the search behavior of MTurk workers using two different methods. In our first method, we scraped pages of available tasks at a very high rate and used the HIT disappearance rates to see if certain sorting methods resulted in greater task acceptance.

Because this method does not accurately measure all types of HITs, we also posted a survey on MTurk asking workers how they search for our survey task. We posted the task with specially chosen posting parameters complementary to those of the HITs we gathered information on using the scraper.

Both methods show that workers tend to sort by *newest* HITs. The scraping revealed that workers also sort by *most* HITs and focus mainly on the first two pages of search results but ignore the positions of the HITs on the page. The survey data confirms that even though *newest* and *most* are the most popular search strategies, the other sort categories are also used—even relatively obscure ones such as *title a-z*

and *oldest* HITs, but to a much lesser extent. The survey confirms that the first two pages of the search results are most important, but also shows evidence that some workers are willing to wade fairly deep into search results pages and to periodically return to the *newest* HITs results.

7. FUTURE WORK

Method A does not account for keyword searching or sorting by required qualifications. The Method A empirical results provide insight into how workers search, but it would useful to develop a structural model of search behavior that incorporated the quantity of work being done could make predictions.

8. ACKNOWLEDGMENTS

John thanks the NSF-IGERT Multidisciplinary Program in Inequality & Social Policy for generous financial support (Grant No. 0333403). Thanks to Marc Grimson for help running the survey and to Greg Little and Jaime Teevan for feedback. Thanks to Robin Y. Horton, Carolyn Yerkes and three anonymous reviewers for very helpful comments. All plots were made using `ggplot2` [18] and all multilevel models were fit using `lme4`[3].

9. REFERENCES

[1] E. Adar, M. Dontcheva, J. Fogarty, and D. Weld. Zoetrope: interacting with the ephemeral web. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 239–248. ACM, 2008.

[2] D. Autor. Wiring the labor market. *Journal of Economic Perspectives*, 15(1):25–40, 2001.

[3] D. Bates and D. Sarkar. lme4: Linear mixed-effects models using S4 classes. *URL http://CRAN. R-project. org/package= lme4, R package version 0.999375-28*, 2008.

[4] S. Brin, L. Page, R. Motwami, and T. Winograd. The PageRank citation ranking: bringing order to the web. In *Proceedings of ASIS'98*, pages 161–172, 1998.

[5] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *The American Economic Review*, pages 242–259, 2007.

[6] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press Cambridge, 2007.

[7] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[8] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K. Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9):49, 2002.

[9] J. Horton and L. Chilton. The labor economics of paid crowdsourcing. *Proceedings of the 11th ACM Conference on Electronic Commerce 2010 (forthcoming)*, 2010.

[10] J. J. Horton, D. Rand, and R. J. Zeckhauser. The Online Laboratory: Conducting Experiments in a Real Labor Market. *NBER Working Paper w15961*, 2010.

[11] B. A. Huberman, D. Romero, and F. Wu. Crowdsourcing, attention and productivity. *Journal of Information Science (in press)*, 2009.

[12] P. Ipeirotis. Demographics of mechanical turk. *New York University Working Paper*, 2010.

[13] E. Law, L. Von Ahn, R. Dannenberg, and M. Crawford. Tagatune: A game for music and sound annotation. In *International Conference on Music Information Retrieval (ISMIR'07)*, pages 361–364. Citeseer, 2003.

[14] W. Mason and D. J. Watts. Financial incentives and the 'performance of crowds'. In *Proc. ACM SIGKDD Workshop on Human Computation (HCOMP)*, 2009.

[15] M. Silberman, J. Ross, L. Irani, and B. Tomlinson. Sellers' problems in human computation markets. 2010.

[16] A. Spink and J. Xu. Selected results from a large study of Web searching: the Excite study. *Inform. Resear.—Int. Electron. J*, 6(1), 2000.

[17] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

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

The Labor Economics of Paid Crowdsourcing

John J. Horton
Harvard University
383 Pforzheimer Mail Center
56 Linnaean Street
Cambridge, MA 02138
john.joseph.horton@gmail.com

Lydia B. Chilton
University of Washington
AC101 Paul G. Allen Center, Box 352350
185 Stevens Way
Seattle, WA 98195-2350
hmslydia@cs.washington.edu

ABSTRACT

We present a model of workers supplying labor to paid crowdsourcing projects. We also introduce a novel method for estimating a worker's *reservation wage*—the key parameter in our labor supply model. We tested our model by presenting experimental subjects with real-effort work scenarios that varied in the offered payment and difficulty. As predicted, subjects worked less when the pay was lower. However, they did not work less when the task was more time-consuming. Interestingly, at least some subjects appear to be "target earners," contrary to the assumptions of the rational model. The strongest evidence for target earning is an observed preference for earning total amounts *evenly divisible by 5*, presumably because these amounts make good targets. Despite its predictive failures, we calibrate our model with data pooled from both experiments. We find that the reservation wages of our sample are approximately log normally distributed, with a median wage of $1.38/hour. We discuss how to use our calibrated model in applications.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics; J.m [**Computer Applications**]: Miscellaneous

General Terms

Human Factors, Economics, Experimentation

Keywords

Crowdsourcing, Amazon's Mechanical Turk, Human Computation

1. INTRODUCTION

Crowdsourcing is a form of "peer production" [2] that outsources work traditionally performed by an employee to an "undefined, generally large group of people in the form of an open call" [11]. Despite the successes and the perceived promise of crowdsourcing,[1] would-be users face a serious practical challenge: they need to attract a crowd. Crowdsourcing projects have used a variety of inducements, including entertainment [21], information [1, 13], and the chance to be altruistic and win attention from others [12]. Until recently, it was extremely difficult to offer money to a crowd, but with the advent of online labor markets such as oDesk, Elance and Amazon's Mechanical Turk (AMT), buyers can now easily pay workers with cash [8].

Compared to cash, non-monetary crowdsourcing incentives are limited in at least two ways. First, some non-monetary incentives depend on the nature of the task or the identity of the proposer, which limits their usefulness. For example, tasks such as classifying web pages, transcribing scanned documents and validating search results are ideal for crowdsourcing, yet they are unlikely to attract volunteers because they are tedious and the private benefits come only to the proposer. Second, even when appropriate non-monetary incentives exist, they are hard to adjust. For example, assuming that we could compute a "fun" labor supply elasticity, ϵ, it is not clear how we could make a task $10(1/\epsilon)\%$ more fun in order to boost provision by 10%. With cash incentives, computing supply elasticities is straightforward and making price adjustments is easy.

The possibility of cash payments raises a design question: how does a would-be user of crowdsourcing effectively and efficiently employ monetary incentives? To solve this problem, designers need two things: (1) a theoretical model that predicts how people will respond to different price/task scenarios and (2) data-driven refinements of that model. The refinements should allow the model to account for behavioral biases, such as those identified by experimental economics, and the idiosyncrasies of particular applications. To borrow an analogy from the economist and market designer Al Roth, just as builders of suspension bridges must be informed by both the elegant theory of mechanics and the messy details of metallurgy and geology, builders of social systems must possess both a general theory of behavior and detailed contextual knowledge [18].

An appropriate theoretical model for the labor supply aspect of crowdsourcing design should tell the designer (1) how workers decide whether or not to participate in a crowdsourcing project and (2) how workers decide the amount to produce, conditional upon participating. In the language of economics, the designer must be able to predict the labor supply on both the (1) extensive and (2) intensive margins.

[1] Examples include Wikipedia, Digg, Yelp, Yahoo! Answers, Stack Overflow and InnoCentive.

EC'10, June 7–11, 2010, Cambridge, Massachusetts, USA.

These two aspects of labor supply have intrigued economists from Adam Smith onward, but research in conventional labor economics is unlikely to directly apply to crowdsourcing "jobs" that last seconds and pay pennies. This caveat aside, labor economics offers a theoretical framework for understanding decision-making in paid crowdsourcing scenarios, and it potentially provides the predictive model that designers need.

1.1 Overview

There are several papers on the design of incentives in crowdsourcing [5] and on obtaining work from AMT [19, 20], but our paper is most closely related to work by Watts and Mason [17], who also conducted labor supply experiments. They found that workers respond to prices in a way that is at least *consistent* with rational behavior. However, their findings also suggest that the relationship between incentives and output is complex: higher pay rates did not improve work quality—a result the authors attribute to high wages affecting worker beliefs. They also found that the structure of incentives (i.e., whether workers faced a piece-rate or a quota system) affected output.

Motivated by the evidence from [17] that AMT workers respond to cash incentives, in Section 2.1 we develop a simple rational model of crowdsourcing labor supply. We also present a novel method for estimating the reservation wage of each worker (Section 2.3) that makes use of a highly concave earnings function rather than a simple piece-rate. The reservation wage is the minimum wage a worker is willing to accept as compensation in exchange for performing some task; it is the key parameter in models of labor supply.

With subjects recruited from AMT, we tested the predictions of the model in two separate experiments. In both experiments, subjects could choose how much output to produce. We tested whether output fell when we made the task more difficult (Experiment A, Section 4) and whether output fell when we lowered the price (Experiment B, Section 5). We find mixed evidence for the rational model: workers are clearly sensitive to price, but they are insensitive to variations in the amount of time it takes to complete a task.

According to the model, reservation wages are fixed and thus should not be affected by our experimental manipulations. Although we find that the imputed wage distributions are quite similar in Experiment A, large differences appear in Experiment B. The cause seems to be that some workers are "target earners" who focus on reaching salient earnings targets. This stands in sharp contrast to the rational model that predicts that workers should consider only the offered wage. These findings have important implications for the design of incentives and are discussed in Section 6.

The rational model clearly misses important elements of reality, but there are no immediate replacements for the model, and it makes reasonable predictions in some cases. For these reasons, we demonstrate in Section 8 how the model can be used to predict labor supply for any price/task scenario, using calibration data pooled from both experiments. We conclude with a discussion of the contributions and limitations of the paper and our thoughts on directions for future research.

2. THEORY

Every time-consuming activity generates an *opportunity cost*. The opportunity cost of doing A is the foregone net benefits one would have obtained from doing a next-best option B. Researchers generally cannot observe the net benefits of a person's options, but when they observe a person doing A, they can infer that the person finds doing A preferable to doing B, with all rewards and costs for the two tasks being taken into account. Applying this inference to work decisions yields a prediction: a person will work only when the net benefits from working exceed the hypothetical net benefits from their next-best alternative, be it another job, leisure or a renewed job search. In labor models, the economic value of this "next-best alternative" is characterized as a reservation wage.

The reservation wage is difficult to estimate in practice for at least two reasons. First, all jobs offer a mixture of non-monetary benefits and costs, or amenities and dis-amenities. For example, there are obvious non-monetary differences between working as a coal miner and working as an ice cream taste-tester. Second, observing someone working tells us only that their total benefits exceed total costs, but not what those total cost actually are. Imagine a job offers a wage w and a stream of amenities, a, and a stream of dis-amenities, d. If the worker works for time t, they receive benefits $(w + a)t$ and bear costs dt. If the worker has a reservation wage, $\underline{\omega}$, then observing someone working tells us only that $w + a - d \geq \underline{\omega}$.

To estimate $\underline{\omega}$, we need to identify the worker's indifference point, i.e., the w^* where $w^* + a - d = \underline{\omega}$. To push a worker down to their indifference point, we could continuously lower their wage by small amounts until they chose to quit. Assuming workers viewed this process sanguinely and their marginal costs were not increasing, their wage when they exit—i.e., when they are indifferent between working and continuing—is their reservation wage for that task. This "decreasing wage" method is clearly impractical in traditional labor relationships, but [4] shows that it can easily be done for small, piece-rate tasks when the process is explained up front and workers have little emotional investment in their seconds-old "job."

In our experiments, we capitalize on this feature of piece-rate work and use a continuously decreasing payment function. To use this method, it is important for workers to be aware that the marginal payment is falling and will continue to fall. Otherwise, incorrect expectations that wages might increase eventually might cause workers to continue working even when wages fall below their reservation wage.

2.1 Model of labor supply

Because the payments in crowdsourcing contexts are so small, we do not expect a worker's marginal utility of wealth to change while working and thus we assume $u(w) = w$. Workers choose some positive, continuous quantity to produce, $y \geq 0$. They are paid $P(y)$ for their choice of y, and we assume that $P(y)$ is strictly increasing, $P'() > 0$, but concave, $P''() \leq 0$. It costs the worker $C(y)$ to complete y tasks. The worker's maximization problem is:

$$\max_y P(y) - C(y) \quad s.t. \quad y \geq 0 \tag{1}$$

The first-order condition is $P'(y^*) = C'(y^*)$, which holds when the marginal benefit of working equals the marginal cost. An interior solution exists only when $P(y) - C(y)$ reaches a global maximum, which occurs when $P(y) - C(y)$ is concave. The concavity of $P(y) - C(y)$ depends jointly on the properties of the payment function and the cost function.

Although corner solutions are interesting and important as predictions of the model, for the purpose of inferring worker reservation wages, we need each worker to choose and an interior solution. We will next discuss how to choose a task and payment function that will lead to this result.

2.2 Cost curves and output

If a task is very tiring, marginal costs are increasing ($C''(y) > 0$), but this is not necessarily the case: costs could be decreasing if workers get better with experience ($C''(y) < 0$), and costs can also be approximately linear ($C''(y) = 0$) or even have different properties at different points (e.g., decreasing at first, then linear).

If $P()$ is linear, then $P(y) - C(y)$ will be concave only when $-C(y)$ is strictly concave, which occurs when marginal costs are increasing. With a constant piece-rate π for each unit of output, the payment function is $P(y) = \pi y$. If costs are not increasing, there is no interior solution and workers are willing to produce either nothing when $\underline{\omega} > \pi/t$ or an infinite amount when $\underline{\omega} < \pi/t$. In the experiments run by [17], workers could choose how many picture-sorting tasks to perform, up to a cap of 100 and with each task paying a constant amount. In the experimenter's high-wage, low-difficulty condition, *mean* output was over 90 tasks, and presumably many of the subjects hit the 100-task cap.

Consistent with our opportunity cost framework, we assume that the only cost of performing a task is the time it takes: $C(y) = \int_0^y \underline{\omega} t(x)dx$, where $t()$ is the marginal completion time and hence $C'(y) = t(y)\underline{\omega}$. The advantage of this assumption is that because the $t(y)$ *curve* is observable, we can determine whether costs are increasing, decreasing or constant. In our experiments, we find that completion times are essentially constant and we assume linear costs.[2] This linearity is unsurprising given the simplicity and brevity of our tasks.

2.3 Using the payment function to impute the reservation wage

Even with linear costs, an interior solution to the maximization problem does not exist unless $P(y)$ is strictly concave. If $P(y)$ is strictly concave, then the maximization problem has an interior solution at $P'(y^*) = \underline{\omega} t_0$. A worker's reservation wage can be estimated directly from their output choice: if they completed y_i^*, then $\hat{\underline{\omega}}_i = \frac{P'(y_i^*)}{\bar{t}_i}$, where $\bar{t}_i$ is the worker's average completion time.

For ease of exposition, we modeled worker output as continuous. In most practical crowdsourcing applications, subjects make discrete output choices. In our experiments, subjects chose some whole number of tasks to complete and $P(y) = \sum_{x=0}^{y} p(x)$, where $p(y) = P(y) - P(y-1)$. When output is discrete, measuring reservation wages is somewhat more complex, but we know that $\underline{\omega}_i \leq p(y_i^*)/\bar{t}_i$ and $\underline{\omega}_i > p(y_i^* + 1)/\bar{t}_i$. If the piece-rate tasks are small and the output choice is fine-grained, then

$$\hat{\underline{\omega}}_i \approx (2\bar{t}_i)^{-1} \left[p(y_i^*) + p(y_i^* + 1)\right]$$

reasonably approximates the reservation wage.

[2]Excluding the very first task, completion times rise by less than 1 *millisecond* per task on average. There is, however, a gap between the first task and second task, with the second task requiring about 1.5 fewer seconds.

3. EXPERIMENTAL PRELIMINARIES

Any reasonable labor supply model should predict that lowering wages will reduce output. There are two ways to lower a person's wages: (a) increase the output they must produce in order to earn their previous wage or (b) lower their wage while keeping the required amount of work constant. As we will show, our model predicts that output will fall in either scenario. We then test these predictions in two experiments. In Experiment A, we test whether increasing the task difficulty reduces output, and in Experiment B, we test whether lowering wages reduces output. Both of our experiments had the same basic set-up: workers from AMT were asked to perform piece-rate tasks. Critically, they were given complete freedom to choose how many pieces to complete. Before we discuss the experiments in depth, we first provide necessary background information.

3.1 Amazon's Mechanical Turk

Amazon's Mechanical Turk is an online labor market where workers can perform "Human Intelligence Tasks" (HITs) for "requesters." HITs vary, but most are small, simple tasks that are difficult for computers but relatively easy for humans. Common tasks include transcribing audio clips, classifying and tagging images, reviewing documents and checking websites for pornographic content. When posting a HIT, a requester describes the task, sets a piece-rate payment, sets worker qualifications, determines how long workers can work on the task, determines how many times they want each HIT performed and creates an interface for workers to use when working on the task.

To become an AMT worker, a person must create an AMT account and provide Amazon with a bank account number. Workers are allowed to have only one account, and Amazon uses several technical and legal means to enforce this restriction. Workers can observe the collection of available HITs and their attributes prior to starting work and can normally view a sample of the required work before starting. They are free to work on any task for which they are qualified, and they can begin work immediately after accepting a HIT. Once a worker completes a HIT, they must submit it for review. A requester then may review the work and decide whether or not to "approve" the HIT. If the HIT is approved, the worker is paid the piece-rate. The worker is also paid if the requester does not review and approve the work within a specified amount of time. Solely at their discretion, requesters may "reject" work, in which case the worker is not paid.[3] Requesters may also elect to pay bonuses, which makes it easy to tailor payments to individual workers based on their performance within a nominally piece-rate HIT.

3.2 Conduct of the experiments

Experiments A and B were conducted in sequence, with approximately 3 days between them. Of the 92 subjects who participated in A, 38 of them also participated in B, which had 198 subjects.

Subjects were not informed that the task had an experimental component. The HIT was simply posted on AMT like any other HIT, with the task title "User interface test." It is important to note that all subjects read identical instructions before accepting the HIT and before observing

[3][10] provides a model of the accept/reject decision in markets like AMT and shows under what conditions an "all reject" equilibrium does not occur.

any unique feature of their assigned treatment group. Thus we are confident that subjects did not select out of the experiment in response to the nature of their assigned treatment. In both experiments, all subjects accepting the HIT submitted a response, so no outcome data are missing.[4]

3.3 Task and user interface

For a crowdsourcing task, we had subjects click back and forth between two narrow vertical bars separated by some number of pixels. The "target" bar was green and the other bar was gray. The target bar alternated between the left bar and the right bar, with the first target bar always beginning on the left. If they missed a bar, the bar flashed red but workers could continue. Figure 1 represents the interface and the cursor movement needed to complete the task.

We chose the clicking task because it is time consuming, requires the full attention of subjects and is not obviously fun. This task is also culturally neutral, easy to understand and easy to make harder (by narrowing the bars or spacing the bars farther apart). We organized work into *blocks*, with each block consisting of 10 back-and-forth clicks. A block is one unit of output (i.e., if a worker completes 3 blocks, $y = 3$). Subjects had 4 seconds to complete each *click*. If they did not perform the click within this time, the HIT ended, but across both experiments, no subject exceeded this cutoff. If a subject missed more than 40% of the clicks in a block, the HIT ended. As with the time cutoff, no subject came anywhere close to this limit. We capped total output at 200 blocks, but this proved unnecessary as the observed maximum output was only 58 blocks.

After completing a block, subjects chose whether to quit or continue. If they had already completed $y - 1$ blocks, they were offered $p(y)$ to perform an additional block (recall that $p(y) = P(y) - P(y-1)$). If they chose to quit, they performed no more tasks and earned $P(y-1)$. When making this exit decision, the interface showed them all the relevant rates and totals, as well as their error rate and their average per-block completion time in seconds. Subjects could rest as long as they liked before starting a new block.

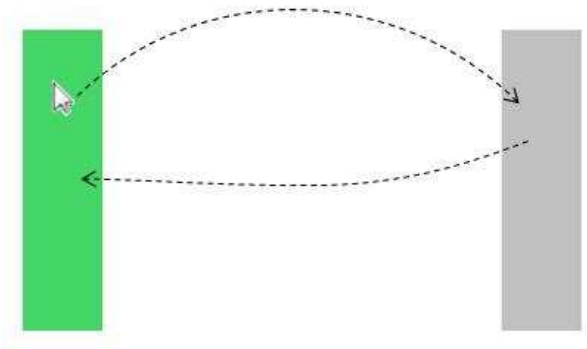


Figure 1: Crowdsourcing task. Subjects were asked to click between the two vertical rectangles. The "target" bar (shown on left here) was always colored green.

[4]A discussion of the causal inference issues raised by revealed-preference experiments conducted online can be found in [9].

3.4 Payment function for our experiments

To create a concave payment function, each subsequent piece of work must pay less than the previous piece of work, but still offer positive payment. One payment function with this property is:

$$P(y) = \bar{P}\left(1 - e^{-ky}\right) \tag{2}$$

Note that total earnings asymptotically approach $\bar{P}$ as y increases. The parameter $k \geq 0$ can be thought of as determining the "half-life" of the payment schedule. For example, if a given k^* has the property that $P(10) = \frac{1}{2}\bar{P}$, then the half-life of the schedule is 10 tasks. In all experiments, we used a half-life of 10; Table 1 shows samples of the total earnings and the marginal payment using Equation 2, given this half-life.[5]

Table 1: Sample earnings and wages with a 10-task half-life

y	$P(y)/\bar{P}$	$p(y+1)/\bar{P}$
1	0.07	0.0625
5	0.29	0.0474
25	0.82	0.0118

One complication in our payment scheme is that it is impossible to pay workers fractional cents. To circumvent this problem, we used a payment system where a worker was paid all of their whole-cent earnings for sure, and their fractional earnings stochastically. If a worker earned h whole cents and f fractions of a cent, then we paid them h with probability $1 - f$, and we paid them $h + 1$ with probability f. Payment was thus correct in expectation, since $\mathbb{E}[P(y)] = (1-f)h + (h+1)f = h + f$. This procedure was explained to subjects before they joined the experiment.

4. EXPERIMENT A: Δ DIFFICULTY

In Experiment A, subjects were randomly assigned to groups *EASY* and *HARD*. The only difference between the two groups was that in *EASY* the vertical bars were 100 pixels apart and in *HARD* the bars were 600 pixels apart. Of the 92 subjects that participated, 42 were assigned to *EASY* and 38 were self-reported females. Subjects completed a total of 18,934 clicks. In both groups, subjects' earnings were determined using the same payment function, Equation 2, with the parameter values of $\bar{P} = 10$ and $k = 10^{-1}\log 1/2$. With these parameter values, total earnings asymptotically approached 10 cents and the "half-life" was 10 blocks, i.e., $P(10) = 5$.

4.1 Model prediction

Using the first-order condition from Equation 1, and assuming a strictly concave payment function and linear costs, $P'(y) = \underline{\omega}t$ (for simplicity, we drop the $*$ notation on y), and treating optimal output as a function of t, we take the total derivative with respect to t and get $y'(t) = \underline{\omega}/P''(y(t))$. We can see that increasing the unit completion time reduces a worker's output because $P''(y(t)) < 0$.

[5]For subjects producing only one unit of output, an adjustment must be made when imputing reservation wages because $P(1)$ needs to incorporate the "show-up" fee.

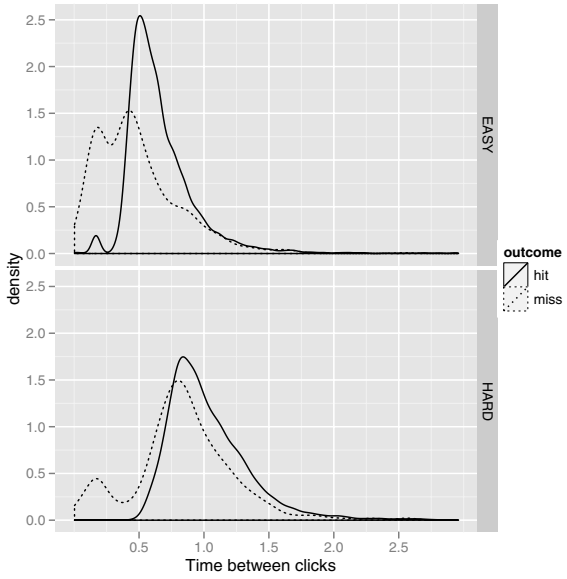


Figure 2: Distribution of time-between-clicks for "hits" and "misses" by treatment group. Distributions are computed using a kernel density estimator.

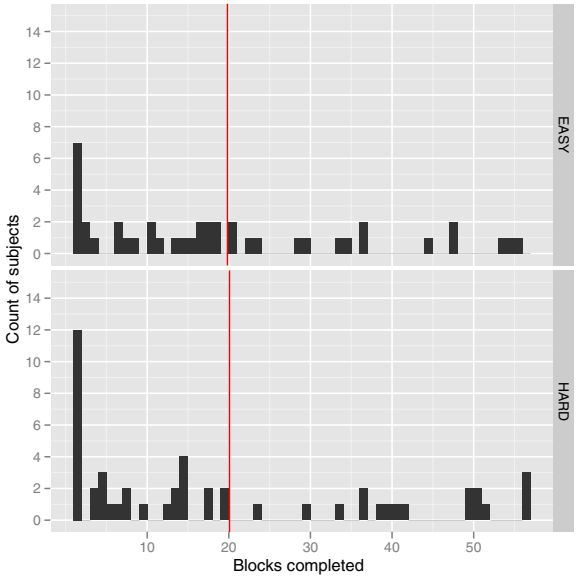


Figure 3: Output by task difficulty. The red vertical line indicates mean output in both groups. Bins have unit length.

4.2 Results

4.2.1 *Effort*

Unsurprisingly, subjects assigned to *HARD* needed more time to complete a block. Regressing average per block completion time (in seconds), $\bar{T}_i$, on the treatment indicator $EASY_i$ (with robust standard errors under each coefficient)[6] we have

$$\bar{T}_i = \underbrace{-4.89}_{1.25} \cdot EASY_i + \underbrace{10.93}_{1.20}$$

with $R^2 = 0.13$ and sample size $N = 92$. The treatment effect is large and highly significant: subjects in *HARD* took about 11 seconds to complete a block; subjects in *EASY* needed only about 6 seconds. If we examine between-click times instead of average block completion times, we see further evidence that the treatment was effective. In Figure 2, the distributions for both "hits" and "misses" are shown for each group. As expected, the "hit" distribution for *HARD* is right-shifted. Confirming our intuitions about the relationship between effort and quality, misses were associated with faster cursor movement.

4.2.2 *Output*

Even though subjects in *HARD* needed more time to complete each block, they had nearly the same pattern of output as subjects in *EASY*. Figure 3 shows output histograms for both groups. There is no discernible difference in mean output, and although more workers in *HARD* quit after performing just one block (12 vs. 7), this difference is not statistically significant.[7]

[6] All standard errors in the paper are robust.
[7] We regressed $1\{y_i = 1\}$ on $EASY_i$.

Regressing output on a group indicator we have

$$y_i = \underbrace{-0.25}_{3.76} \cdot EASY_i + \underbrace{20.08}_{2.72}$$

with $R^2 = 5e-05$ and the sample size $N = 92$. Subjects in *EASY* had slightly less average output, though the effect is imprecisely estimated. Transforming output with the log function and running the same regression, we have

$$\log y_i = \underbrace{0.14}_{0.28} \cdot EASY_i + \underbrace{2.30}_{0.20}$$

with $R^2 = 0.00256$ and $N = 92$. Even though the coefficient on *EASY* changes signs, both the regressions and the graphical analysis of Figure 3 lead to the same conclusion: despite the greater time required per block for subjects in *HARD*, there was no discernible difference in the patterns of output across groups.

4.2.3 *Reservation wages*

We imputed the reservation wage for each subject using the method explained in Section 2.3. The distribution of the log of estimates is plotted in Figure 4, which shows both the smooth kernel density estimate of the distributions as well as the the actual estimated values (displayed as tick marks along the horizontal axis). The top two panels contain the results of Experiment A. We can see that the imputed reservation wage distributions are quite similar, except that *HARD* has a fat left tail, implying that a cluster of workers in *HARD* exhibited output patterns consistent with very low reservation wages. However, this could be due to sampling variance. Indeed, in a regression of log reservation wages on a group indicator we have

$$\log \hat{\omega}_i = \underbrace{0.52}_{0.37} \cdot EASY_i + \underbrace{-0.12}_{0.26}$$

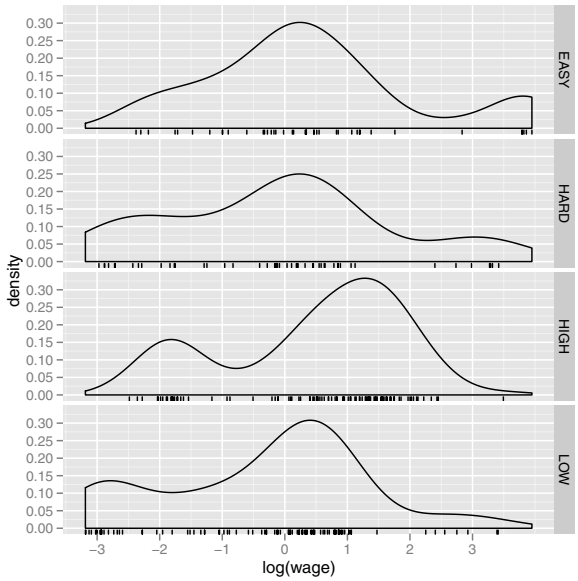


Figure 4: Imputed log reservation wage distributions both experimental groups, in both experiments.

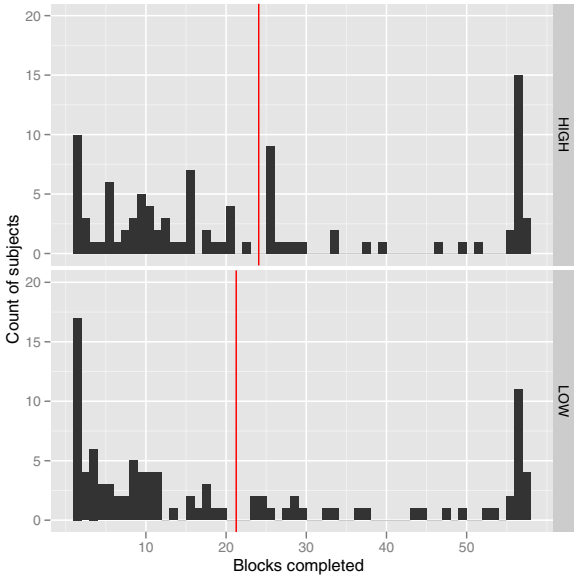


Figure 5: Output by price group. The red vertical line indicates mean output in both groups. Bins have unit length.

with $R^2 = 0.02$ and sample size $N = 92$. Assignment to *EASY* had a positive but statistically insignificant effect. The coefficients in this regression are more easily interpretable following transformation. They imply that the geometric mean reservation wage was \$0.89/hour in *HARD* and \$1.49/hour in *EASY*.

4.3 Discussion

The experimental results are theoretically ambiguous but are internally consistent. Given that workers in *HARD* took longer per task but had essentially the same output pattern as subjects in *EASY*, we would expect subjects in *HARD* to have relatively higher reservation wages, which is what we find, albeit the effect has a t-statistic of only 1.41.

Perhaps the simplest explanation for the lack of treatment effects is that our treatment was not strong enough and that workers are not attuned to fairly small differences in time. Lord Robbins' notion of the "marginal utility of not bothering with marginal utility" seems particular apt, especially if workers have a heuristic task-based (as opposed to time-based) reservation wage [16].

5. EXPERIMENT B: Δ PRICE

In Experiment B, 198 subjects were randomly assigned to groups *HIGH* and *LOW*. The task was identical in both groups, with the bars 100 pixels apart, but earnings asymptotically approached either 10 cents (in *LOW*) or 30 cents (in *HIGH*). Because of the imprecise treatment effect estimates in Experiment A, we doubled the sample size. By chance, both groups had the same number of subjects. Of the 198 subjects, 72 were self-reported females. Subjects completed a total of 45,710 clicks.

5.1 Model prediction

Consider an alternative payment function $\gamma P(y)$, where γ is some positive number. The first-order condition for the original maximization problem is now $\gamma p(y) - t\underline{\omega} = 0$, and if we treat the optimal output y as a function of γ, the total derivative of the first-order condition with respect to γ is $p(y) + \gamma p'(y(\gamma))y'(\gamma) = 0$ and thus:

$$y'(\gamma) = -\frac{p(y)}{\gamma p'(y(\gamma))}$$

Because payment is strictly increasing, $p() > 0$, and because the total payment function is concave, $p'() < 0$, it follows that $y'(\gamma) > 0$, i.e., output increases when payment is higher.

5.2 Results

5.2.1 Output

Figure 5 shows the histogram of output in the two groups. Although *LOW* has more early quits, both groups have sizable numbers of early quitters as well as some stalwarts that completed in excess of 50 blocks. Unlike in Experiment A, mean output in *LOW* is noticeably lower than in *HIGH*. Regressing output on a group indicator we have

$$y_i = \underbrace{-2.81}_{2.89} \cdot LOW_i + \underbrace{24.07}_{2.00}$$

with $R^2 = 0.0048$ and $N = 198$. Subjects in *LOW* had lower mean output, but the effect is imprecisely estimated due to the bi-modality of output, which can be seen in the output histograms. Estimating the regression in logs rather than levels we have

$$\log y_i = \underbrace{-0.30}_{0.17} \cdot LOW_i + \underbrace{2.71}_{0.11}$$

with $R^2 = 0.02$ and $N = 198$. There is a large and significant difference in geometric means across the groups. From Figure 5, it appears that most of the effect is due to the relatively large number of subjects in *LOW* quitting after small amounts of output. We can confirm this graphical observation by regressing an indicator for whether a subject completed fewer than 10 blocks on the treatment indicator:

$$1\{y_i < 10\} = \underbrace{0.152}_{0.07} \cdot LOW_i + \underbrace{0.273}_{0.05}$$

with $R^2 = 0.03$ and $N = 198$. The choice of 10 blocks is somewhat arbitrary, but the effect is strong for all low cutoff values, and a QQ-plot (not shown) makes the pattern even more apparent.

5.2.2 Reservation Wages

Unlike in Experiment A, group assignment strongly affected the imputed reservation wage. The bottom two panels of Figure 4 show that the mass of observations at the center of the distribution is left-shifted for *LOW* relative to *HIGH*. Both distributions also have a second mass of observations with very low reservation wages, but as with the main hump, the *LOW* observations are shifted further left than the low-wage hump in *HIGH*. Regressing log wages on a group indicator we have

$$\log \hat{\omega}_i = \underbrace{-0.79}_{0.22} \cdot LOW_i + \underbrace{0.45}_{0.14}$$

with $R^2 = 0.06$ and $N = 198$. Transforming the predictions into levels, the geometric mean reservation wage was \$1.57/hour in *HIGH* and \$0.71/hour in *LOW*.

5.3 Discussion

As per the model's prediction, lower pay reduces output, but the finding that being in *LOW* lowers a worker's reservation wage implies that the lower output in *LOW* is not as low as it *should* be. Because the actual tasks are identical across groups, the difference in imputed reservation wages cannot be due to uncontrolled-for differences in amenities. There are several other possible explanations for the data, including worker error, increasing but non-time-based marginal costs and target earning behavior; we believe target earning provides the most compelling explanation for our findings.

Worker error could explain our results: we would find lower reservation wages in *LOW* if (1) subjects in *LOW* falsely believe that they are being paid more than they actually are, (2) subjects in *HIGH* falsely believe that they are being paid less than they actually are or (3) some combination of (1) and (2). Liebman and Zeckhauser [15] argue that people systematically misinterpret schedules, usually by letting early marginal payments "bleed over" to affect perceptions of future marginal payments. However, with our concave schedules, this bleed-over would occur in both groups. While we have no evidence on this question, it seems likely that any bleed-over effects would be greater in *HIGH*, which would lead to effects in the opposite direction of what we do find. Furthermore, because subjects were informed of their precise marginal payment after each block, this bleed-over explanation seems unlikely.

Constant marginal costs are a key assumption in our reservation wage estimation method. If marginal costs are rapidly increasing, then we would incorrectly infer that subjects in *LOW* had reservation wages too low: the additional output that subjects in *HIGH* should produce because of their higher wages is moderated by the higher costs of high output. Although an explanation involving increasing marginal cost is possible, it seems unlikely, given that subjects could rest for as much time as they liked between blocks. Furthermore, even subjects producing lots of output spent less than 15 minutes total, including resting time (recall that per block time for the 100-pixel group was less than 7 seconds). Rapid increases in marginal costs over such a short time seem rather unlikely.[8]

An additional explanation for the results is that some workers may be *target earners*. Target earners try to obtain some self-imposed earnings goal rather than respond to the current offered wage. If at least some workers are target earners, then the results are easy to explain: because their wages are lower, subjects in *LOW* must produce more output to achieve their earnings targets than they would have to produce if they were instead assigned to *HIGH*.

6. DEPARTURES FROM THE RATIONAL MODEL

The differences in the imputed reservation wage distribution—particularly those found in Experiment B—strongly suggest that the rational model cannot explain the behavior of all workers. Our best conjecture is that workers create earnings targets that influence their output decisions. While having goals seems sensible and perhaps heuristically "rational" if the target provides motivation, earnings targets lead workers to commit a kind of sunk-cost fallacy, because in the absence of income effects, past earnings are irrelevant to the decision they must make at the margin (i.e., whether the next bit of earnings is worth the the trouble of the next bit of work).

To return to the analogy of suspension bridge building from Section 1, we are now dealing with metallurgy instead of mechanics: target earners do not behave like rational workers in certain contexts, and this difference will matter in applications. For example, in the absence of income effects, when wages are high, a target earner works less and a rational worker works more. There is mixed evidence on the question of whether target earning occurs in "real life" [3, 6, 7], but we find fairly unambiguous evidence of target earning in several places in the results.

The first piece of evidence of target earning is that in every experiment, at least some subjects try to pursue the maximum earnings possible, despite the low wages associated with this strategy. For rational workers to generate this pattern, the wage distribution would have to be highly bi-modal. A more plausible explanation is that workers try to earn the full amount possible (i.e., $\bar{P}$ is their target) and quit only when they realize this goal is unattainable.[9]

[8] One possibility we must consider with more complex work is that workers are likely to need more experience with a task before deciding whether they are good at it. In these cases, we might mistakenly view learning and exit as increasing marginal costs.

[9] A more pedestrian explanation could be that some subjects believe that the payment is in dollars, not cents, despite the clear instructions that stated that payment would be in cents. One subject did email us insisting that payment was supposed to be in dollars—we sent him screen shots proving this was not the case.

6.1 Preferences for "focal point" earnings

More evidence of target earning can be see in the pattern of output. In Figure 6, we plot histograms of the output for *HIGH*, grouped in horizontal panels by the floor of earnings, $\lfloor P(y) \rfloor$. Subjects show a preference for working the minimum amount possible to earn some *whole* number of cents: when multiple output values yield the same number of whole cents, the far left bar of the histogram is the highest. The only exception is in the 29-cent panel; however, recall that in *HIGH*, subjects' earnings asymptotically approached 30 cents. Subjects quit once they realized that they could not break out of the 29-cent band. Although suggestive of targeting, the preference for whole cents could be symptomatic of extreme risk aversion (recall from Section 3.2 that payment is stochastic due to the fractional cent problem), or workers could believe that we might cheat them on the fractional cents (i.e., ignore their fractional earnings) but that we would not be willing to withhold whole-cent earnings.

Harder to reconcile with a rational model is another pattern seen in Figure 6: subjects show a preference for earnings amounts evenly divisible by 5. The smallest earnings amounts (e.g., 2, 3 and 5 cents) all get several subjects, but because these are people who quit very early, they presumably do not have a target or would not need a target. However, we see clear output spikes at 15, 20 and 25 cents. Assume that in the absence of "modulo 5" targeting, the proportion of subjects earning amounts divisible by 5 should be equal, in expectation, to the proportion of potentially realizable amounts divisible by 5. Consider the set of possible whole-cent earnings:

$$\mathbf{P_w} = \{\lfloor P(1) \rfloor, \lfloor P(2) \rfloor, \ldots \lfloor P(N) \rfloor\}$$

and the fraction of those elements of $\mathbf{P_w}$ that are divisible by 5, which is given by:

$$q = |\mathbf{P_w}|^{-1} \sum_{x \in \mathbf{P_w}} 1\{x \mod 5 = 0\}$$

where $|.|$ indicates the number of elements in the set. Let $\mathbf{Y}$ be the set of realized output choices for an experiment group and let n be the number of observations, $n = |\mathbf{Y}|$, and let s be the actual number of observations in Y such that the whole-cent earnings were divisible by 5: $s = \sum_{y \in \mathbf{Y}} 1\{\lfloor P(y) \rfloor \mod 5 = 0\}$.

Under our assumption of proportionality, we can compute the probability of observing s successes out of n trials when the per-trial probability of success is q. We had 33 successes out of 99 trials when the probability was only $q = 0.22$. The probability of observing this many successes or more by chance is only 0.0027. If we exclude subjects in the 29-cent band, the result is even more compelling: the probability is only $2.55e-05$.

7. USING THE CALIBRATED MODEL

It is straightforward to use our calibrated model for prediction. Consider some crowdsourcing task that takes t_0 to complete and that will pay a piece-rate of p_0. Workers whose reservation wage exceeds the offered wage, p_0/t_0, will accept the task. The fraction of workers producing at least one unit of output is equal to the probability that a given worker will find participating attractive: $Pr(p_0 \geq \underline{\omega}_i t_0) = F\left(\frac{p_0}{t_0}\right)$, where F is the cumulative density function of the estimated reservation wage distribution.

The labor supply curve is $S(w) = N_s F(\log w)$, where N_s is the number of workers in the population and $w = p_0/t_0$. The point elasticity of extensive labor supply is thus $\epsilon_w = f(\log w)/F(\log w)$. To compute the intensive elasticity, we would have to know something about the change in marginal costs.

If we pool estimates from both experiments and assume that wages are log normally distributed, the distribution parameters are $\hat{\mu} = 0.074$ and $\hat{\sigma} = 1.634$, with the reservation wage measured in dollars per hour. In Table 2, the 25th, 50th and 75th percentiles of the pooled reservation wage distribution are shown, as well as the extensive labor supply elasticity computed at that point using the log normal approximation. Consistent with a right-skewed distribution like the log normal, the arithmetic mean is considerably higher than the median wage.

Table 2: Pooled reservation wage distribution properties

	Wage ($/hour)	Point Elasticity
25th p.	0.321	0.81
Median	1.384	0.43
75th p.	2.876	0.28
Mean	3.625	0.24

It is important to remember the fairly strong assumptions underlying these predictions. First, the reservation wage distribution was estimated using a selected sample of AMT workers willing to participate in our experiment. Second, we are assuming constant marginal costs, and we are assuming that task amenities/dis-amenities for the new task are equivalent to those from the back-and-forth clicking task we used. Third, we are assuming that there is a fixed reservation wage and that workers respond rationally to the offered wage, despite some evidence to the contrary.

8. CONCLUSION

We find some agreement with a simple rational model, as well as important anomalies; we find fairly strong evidence that at least some workers work to targets. Designers should consider this propensity when designing incentives schemes and give people natural targets that will increase output, though they should also consider that such schemes might seem manipulative and could backfire (and potentially be unethical).

In the following section, we demonstrate how our calibrated model can be used in applied work. While we find only partial agreement between the model's predictions and reality, we believe it still offers a useful approximation. We conclude by discussing how our paper fits into a larger research program and laying out some directions for future investigations.

8.1 Future research

Crowdsourcing is still a new development, and many open questions remain. Perhaps the most obvious next empirical step is to collect data on the compensating differentials associated with different kinds of crowdsourcing tasks. In other words, compared to a baseline task, how much more or less do workers have to be paid to generate the same amount of output? A related question is how costs change with output,

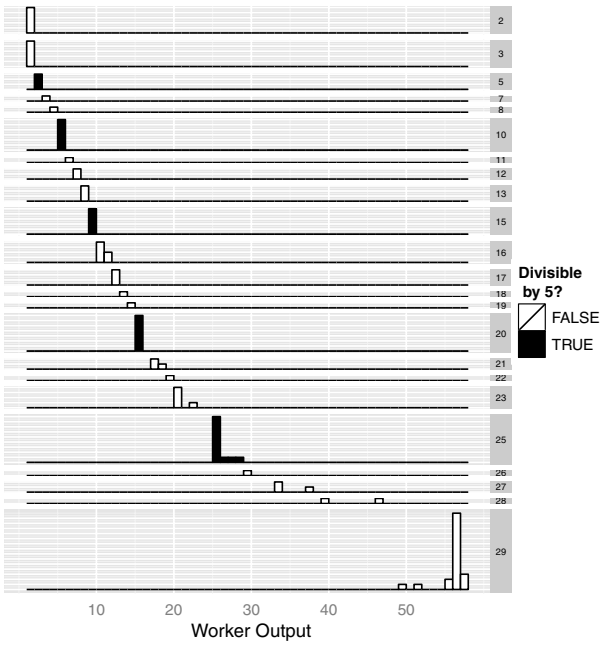


Figure 6: Earnings in the *HIGH* Group. This plot shows the histogram of output, grouped by the whole-cent portion of worker earnings (in cents).

i.e., does a task get much easier or much harder as the worker gains experience? Finally, it would be interesting to learn about the correlates of reservation wages. For example, do groups with lower opportunity costs (e.g., the unemployed, non-US citizens, etc.) have lower reservation wages?

Perhaps the biggest limitation of the current model is that it only predicts the fraction of workers that will accept a task. It tells us nothing about how many workers will actually see a given offer on AMT. Furthermore, AMT workers are not uniformly distributed across time zones, and presumably the number of potential workers waxes and wanes over the day. Understanding search and job uptake would be useful for applications—search is a very important and yet comparatively under-studied topic in economics.

8.2 Research approach to crowdsourcing

We view our paper as a step towards the theoretical framework for crowdsourcing/human computation advocated by [14]. Jain and Parkes argue (correctly in our view) that advances in crowdsourcing as a methodology will require broadly applicable, predictive models. They discuss game theory as a potential generalizing framework, but point out that crowdsourcing "games" are not so much about workers revealing private information (which would suggest a mechanism design approach) as they are about getting workers to show up and exert the effort needed to accomplish a task.

In a labor relationship, there is a "game" between workers and their employers: workers must choose how much (costly) effort to provide, and employers—who prefer high effort—cannot directly observe this choice. This classic principal/agent problem arises in many real-world contexts, but we argue that in *most* crowdsourcing applications, it is a fairly easy problem to solve.[10] When effort is highly correlated with output and output is observable, simple mechanisms can eliminate moral hazard. By rejecting low-quality work (or not hiring low-quality workers in the future), buyers can easily make shirking a dominated strategy.

Although the moral hazard problem is solvable, the problem of getting a sufficient supply of labor persists. Once an employer sets up a quality-control mechanism (e.g., screening, firing, spot checks, etc.), the worker is essentially playing a game against nature. Individual workers will make labor supply decisions by comparing the costs and benefits of working, and although workers must think rationally about their preferences, they do not have to think strategically. In short, we do not need a game theory of crowdsourcing, but rather a *price theory* of crowdsourcing. Our model and reservation wage estimation procedure provide the groundwork for such a theory.

9. ACKNOWLEDGMENTS

John Horton thanks the Berkman Center for Internet and Society and the NSF-IGERT Multidisciplinary Program in Inequality & Social Policy for generous financial support (Grant No. 0333403). Thanks to Richard Zeckhauser for extremely useful comments and suggestions. All plots were made using the open source R package `ggplot2`, developed by Hadley Wickham [22]. The experimental interface was developed using Adobe Flex Builder 3, data was captured using MySQL and the experiment was run on an Ubuntu 8.04.2 LTS (Hardy Heron) server.

[10]We qualify this statement because problems do arise when buyers do not know "what right looks like" and aggregating the inputs of other workers is not helpful for determining quality.

10. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo Answers: Everyone knows something. In *Proc. 17th Intl. Conf. on the World Wide Web (WWW)*, 2008.

[2] Y. Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2007.

[3] C. Camerer, L. Babcock, G. Loewenstein, and R. Thaler. Labor supply of New York City cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112(2):407–441, 1997.

[4] D. L. Chen and J. J. Horton. The wages of paycuts: Evidence from a field experiment. *Working Paper*, 2009.

[5] D. DiPalantino and M. Vojnovic. Crowdsourcing and all-pay auctions. In *Proc. 10th ACM Conf. on Electronic Commerce (EC)*, 2009.

[6] H. S. Farber. Reference-dependent preferences and labor supply: The case of New York City taxi drivers. *American Economic Review*, 98(3):1069–1082, 2008.

[7] E. Fehr and L. Goette. Do workers work more if wages are high?: Evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317, 2007.

[8] B. Frei. Paid crowdsourcing: Current state & progress toward mainstream business use. *Whitepaper Produced by Smartsheet.com*, 2009.

[9] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory. *Working Paper*, 2010.

[10] J. J. Horton and X. Zhu. The allocation of decision rights in designed markets. *Working Paper*, 2010.

[11] J. Howe. *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*. Three Rivers Press, 2009.

[12] B. A. Huberman, D. M. Romero, and F. Wu. Crowdsourcing, attention and productivity. *J. Inf. Sci.*, 35(6):758–765, 2009.

[13] S. Jain, Y. Chen, and D. C. Parkes. Designing incentives for online question and answer forums. In *Proc. 10th ACM Conf. on Electronic Commerce (EC)*, pages 129–138, 2009.

[14] S. Jain and D. C. Parkes. The role of game theory in human computation systems. In *Proc. ACM SIGKDD Workshop on Human Computation (HCOMP)*, 2009.

[15] J. Liebman and R. Zeckhauser. Schmeduling. *Working Paper*, 2004.

[16] R. Lionel. *An Essay on the Nature and Significance of Economic Science*. London, Macmillan, 1935.

[17] W. Mason and D. J. Watts. Financial incentives and the 'performance of crowds'. In *Proc. ACM SIGKDD Workshop on Human Computation (HCOMP)*, 2009.

[18] A. E. Roth. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70:1341–1378, 2002.

[19] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label?: Improving data quality and data mining using multiple, noisy labelers. In *Proc. 14th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2008.

[20] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[21] L. Von Ahn. Games with a purpose. *IEEE Computer Magazine*, 39(6):96–98, 2006.

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

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

John J. Horton
Harvard University*

August 27, 2011

Abstract

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

JEL J01, J24, J3

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

1 Introduction

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

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

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

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

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

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

1.1 Overview of experiments and findings

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

1.2 Interpretation

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

2 Conceptual framework

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

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

2.1 Social learning does not necessarily promote homogeneity

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

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

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

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

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

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

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

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

2.1.1 Evaluation and social preference

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

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

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

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

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

2.2 Related work of peer effects in employment settings

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

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

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

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

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

3 Methods and Materials

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

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

Table 1: Details of the Experiments

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

3.1 Amazon Mechanical Turk

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

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

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

3.2 Task and interface

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

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

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

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

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

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

4 Experiment A: Perceived employer expectations

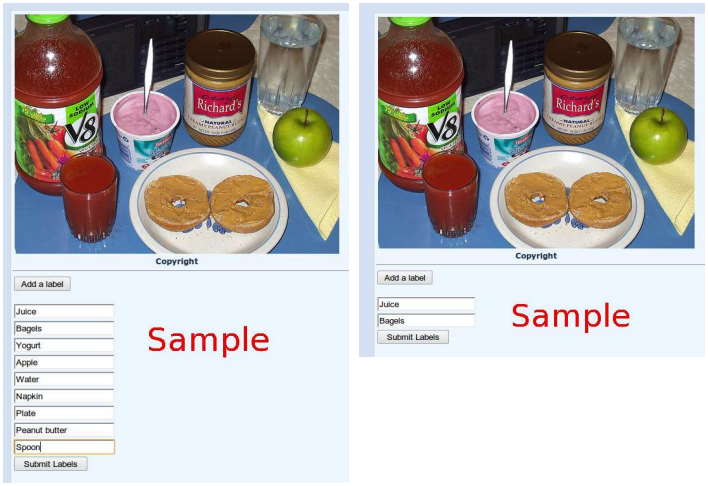
Experiment A tested whether a firm in this market can influence worker productivity by manipulating worker perceptions of employer expectations. Would-be participants were assigned to one of two experimental groups: $HIGH$, in which the work sample

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

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

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

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



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

4.1 Results

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

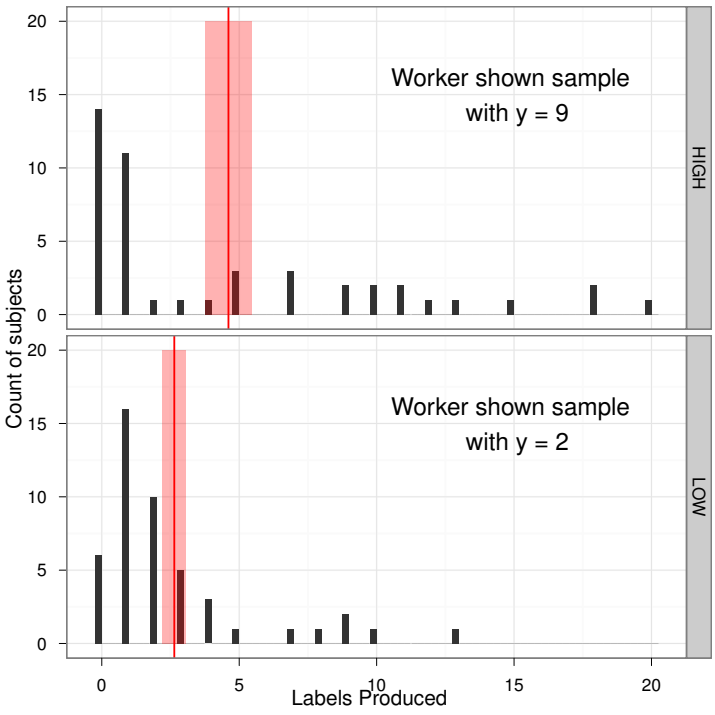


Figure 2: Histogram of labels produced by experimental group in Experiment A. The solid vertical line indicates the mean, while the shaded band around that line is 2 standard errors wide. Subjects in the $HIGH$ group were shown a work sample consisting of 9 labels prior to performing, while subjects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

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

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

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

4.1.1 High employer expectations reduced labor supply on the extensive margin

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

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

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

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

4.1.2 High employer expectations increased labor supply on the intensive margin

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

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

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

4.2 Discussion

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

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

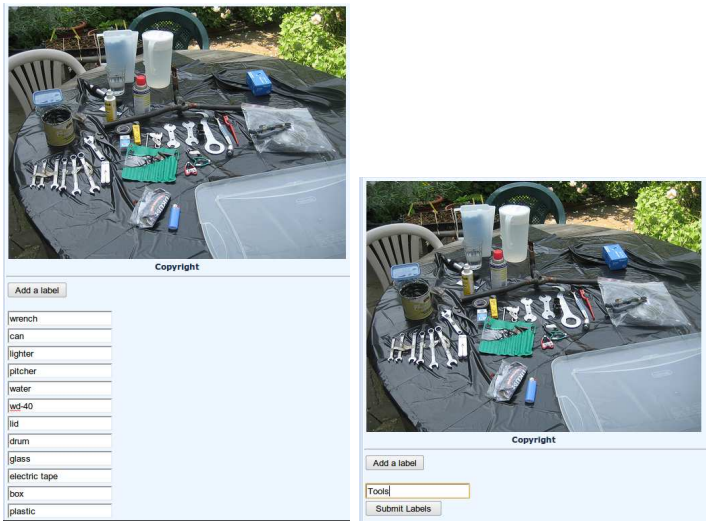
5 Experiment B: Punishment and peer output

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

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

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

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



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

Figure 3: Work evaluated by subjects in Experiment B

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

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

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

5.1 Results

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

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

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

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

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

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

5.1.2 Workers rewarded better work with more generous bonuses

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

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

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

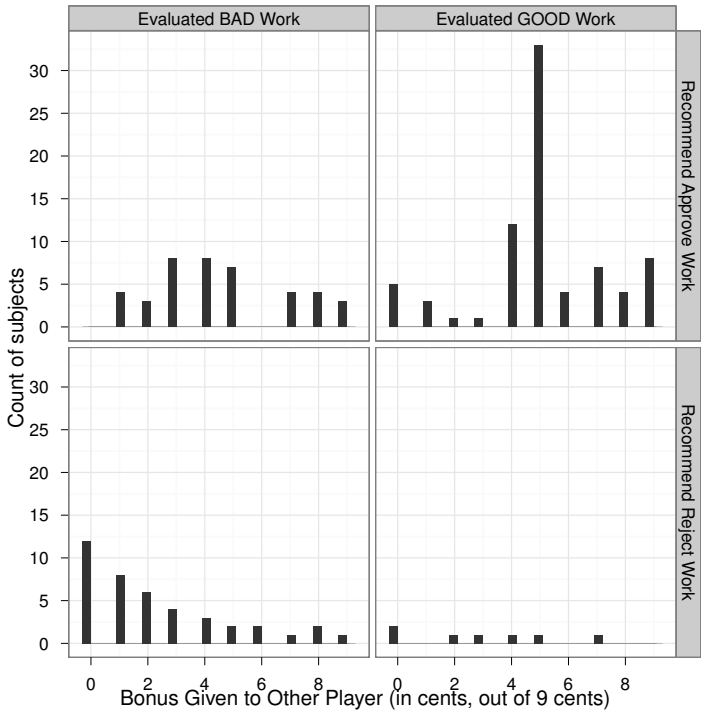


Figure 4: Experiment B, bonus allocation by treatment group and accept/reject recommendation. Subjects in *BAD* evaluated work with 1 generic label, while subjects in *GOOD* evaluated work with 12 specific and appropriate labels.

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

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

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

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

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

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

5.2 Discussion

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

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

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

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

6 Experiment C: Productivity and punishment

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

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

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

6.1 Pilot experiments

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

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

6.2 Actual experiment

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

6.3 Results

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

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

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

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

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

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

6.3.1 Workers with reduced output more generous

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

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

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

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

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

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

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

6.3.2 Workers with reduced output less likely to punish

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

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

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

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

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

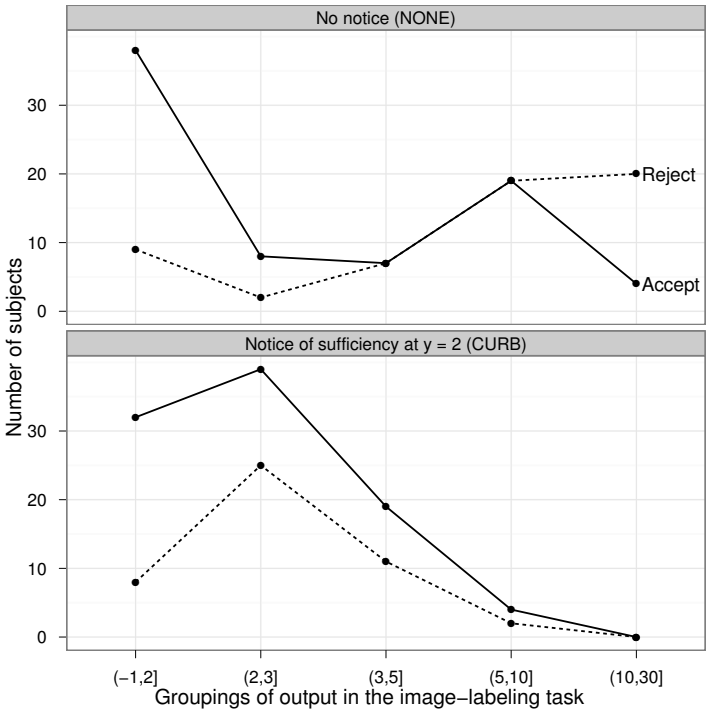


Figure 5: Numbers of subjects recommending either acceptance or rejection at different output levels (indicated by line types), faceted by experimental groups, in Experiment C . Note that a disproportionate number of subjects not receiving the output-curtailing message (subjects in $NONE$) produced relatively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

6.4 Discussion

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

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

7 Experiment D: Peer effects from evaluation

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

7.1 Results

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

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

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

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

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

7.1.1 Exposure to highly productive peers increased productivity

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

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

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

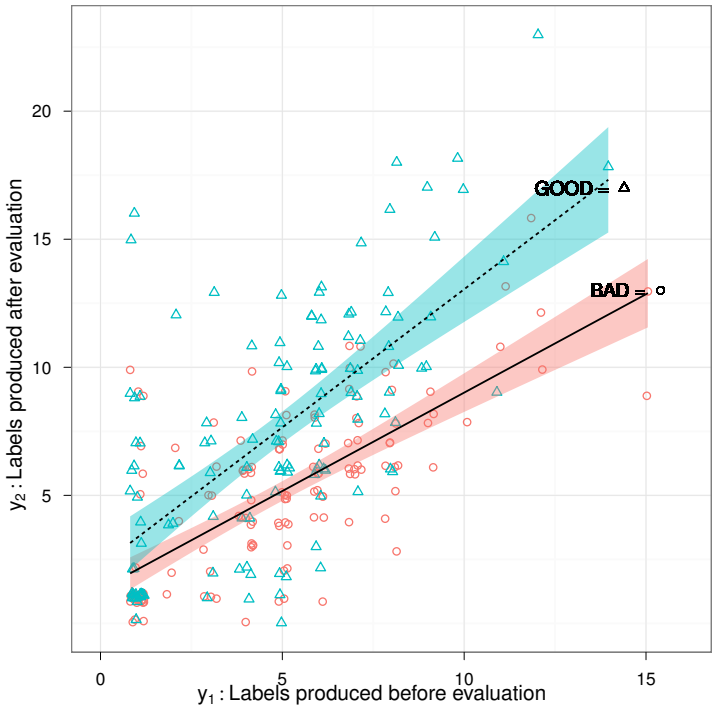


Figure 6: Follow-on output, y_2, versus initial output,y_1, in Experiment D, by group. All subjects did an identical initial task and chose some number of labels to provide (shown on the x-axis). Subjects then evaluated another subject's work that demonstrated either low productivity (BAD) or high productivity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

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

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

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

7.2 Discussion

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

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

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

8 Experiment E: Peer effects after explicit employer instructions

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

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

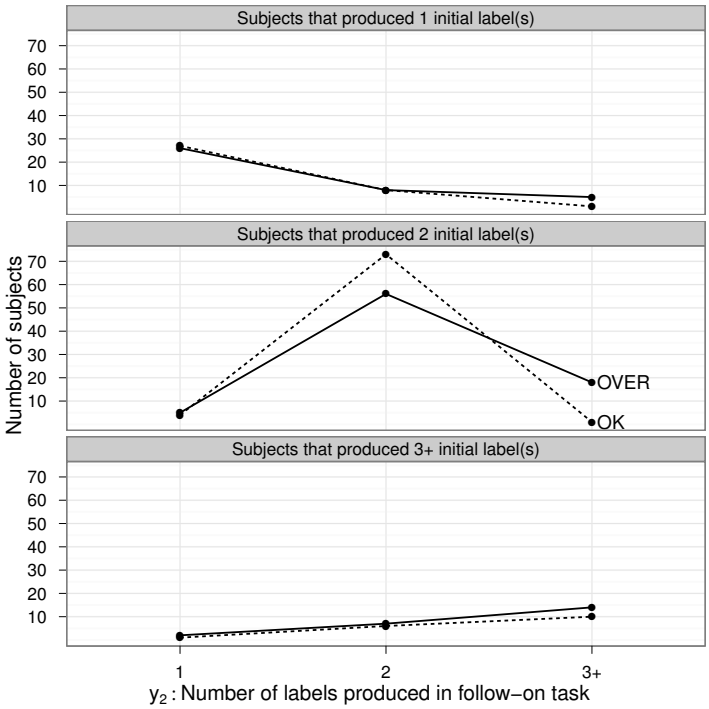


Figure 7: Numbers of workers in different output bands for the follow-on task in Experiment E, faceted by initial output, with experimental groups indicated by line type . Note that for initially complying subjects (middle panel), assignment to $OVER$ increased the relative number of subjects producing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

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

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

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

8.1 Results

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

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

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

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

8.1.1 Most workers compliant with employer output requests

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

8.1.2 Language barriers likely prevented full initial compliance

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

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

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

8.1.3 Evaluating compliant work increased compliance for initially complying workers

Being assigned to OK strongly increased y_2-compliance:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

8.2 Discussion

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

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

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

9 Conclusion

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The Wages of Pay Cuts: Evidence from a Field Experiment*

Daniel L. Chen
Harvard Law School
dlchen@post.harvard.edu
http://www.people.fas.harvard.edu/~dlchen

John J. Horton
Harvard Kennedy School
horton@fas.harvard.edu
http://www.people.fas.harvard.edu/~horton

Abstract

To understand why firms rarely cut nominal wages, we hired workers for a data entry task, paid them a high wage and then offered some of the workers the opportunity to keep working, albeit for a lower wage. We framed the new wage offer in different ways across treatment groups. Workers were more likely to reject lower offers, but "reasonable" justifications largely eliminated this effect. Not all justifications were effective—justifying the cut on the ground that it would increase our profits actually increased quits. We also measured whether the treatments affected quality, trust and cooperation. The "profits" treatment reduced cooperation and possibly reduced quality; the other treatments had generally weak or nonexistent effects.

JEL Codes: J30, D03, D63

This Draft: 2 April 2009

1 Introduction

Firms rarely cut nominal wages, even during recessions. If they must reduce their wage bill, employers generally reduce hours or lay off workers. Because this pattern of downward nominal wage rigidity has import consequences for policy and for our understanding of labor markets, economists have proposed several explanations, but

*The authors thank Nolan Miller, Jeffrey Liebman and Richard Zeckhauser for helpful discussions. Work on this project was conducted while Daniel Chen received financial support from the Institute for Humane Studies, the John M. Olin Foundation, and the Petrie-Flom Center. John Horton received financial support from the Harvard Multidisciplinary Program in Inequality & Social Policy. Both authors gratefully acknowledge funding from the Berkman Center for Internet and Society at Harvard Law School. This is a preliminary draft, but please feel free to distribute—comments are welcome.

empirical researchers have found discriminating among these theories difficult. The difficulty stems in part from a lack of exogenous variation in firm wage setting and in part because of the inherent difficulty of explaining with data why something does not happen.[1] In response to this empirical challenge, Bewley (1999) took the atypical step of interviewing a large number of firm managers and directly asking them how they determined their wage policies, particularly why they did not cut wages during recessions. The managers reported that cuts would hurt morale, lower workers' identification with the firm and ultimately reduce profitability. The usual caveats about self-reports aside, this explanation for the absence of cuts—fear of how workers will react—seems plausible, and yet it raises another question, which is why workers would react so badly to wage cuts.

We can, of course, understand why workers are not *happy* with lower wages, but inflation effectively lowers wages constantly and firms often let real wages fall. Despite having the same effects on purchasing power, there are two important and obvious difference between nominal wage cuts and real wage erosion: wage cuts are highly salient, and people generally view acts of omission as less blameworthy than than acts of commission. This saliency distinction suggests a mechanism for nominal wage rigidity: if workers (a) appreciate that wages are being cut and (b) view those cuts as unfair and (c) are willing to altruistically punish perceived wrongdoers, firms might find that the costs of cutting wages outweigh the benefits.

The idea that nominal wage rigidity is explained by firms setting wages in the shadow of worker fairness reactions seem plausible, but it is unsatisfying, or least not that helpful, because it hinges on what workers are likely to *perceive* as fair. Questions about perception take us far from the standard purview of economics, and at least according to some psychologists, fairness judgments can be highly context-dependent. Given the impossibility of cataloging—never mind experimentally manipulating—every possible context in which wage setting occurs, we are left in the unfortunate state of knowing that an explanatory factor is both important and hard to model or generalize.

In this paper, we attempt to sidestep the multiplicity of contexts problem by assuming that even though there are many context-dependent paths to the verdict of "fair" or "unfair," worker behavior—conditional upon reaching one judgment or the other—is predictable. To implement this idea, we conducted an experiment in which we hired workers for a data entry task, paid them a high wage and then offered

[1] Theoretically, we know why people do not burn $100 bills, but we do not have much observational evidence on the subject.

"treated" workers the opportunity to keep working, albeit for a low wage. Critically, this low wage offer was framed or justified in different ways across treatment groups. We designed the different framings and justifications so that they would provoke workers to view the wage cuts as either reasonable and fair or capricious and unjust. After offering the wage cut, we observed whether workers accepted our offer and we observed how workers responded along several dimensions firms are likely to care about, such as work quality, cooperation with the firm, trust in the firm and patience.

We found that workers were more likely to reject low offers, but "reasonable" justifications largely nullified this effect. Not all justifications were effective—justifying the cut in terms of our profits actually increased quits. This "profits" treatment also reduced cooperation and quality. The other treatments had generally weak or nonexistent effects and none of the treatments affected trust or patience.

2 Theory & evidence

There are a multitude of possible reasons why wages might affect productivity and hence there are a multitude of "efficiency wage" models.[2] One sub-family comprises models that focus on the relationship between pay and worker effort. For example, in the "shirking model," firms pay above-market wages to gain leverage over workers; an efficiency wage job is valuable, and a worker might refrain from shirking to lower the risk of being fired (Shapiro and Stiglitz, 1984). In the "gift exchange" or the closely related "fair wage" models, firms give a "gift" of high wages and workers respond by gifting consummate performance; if workers feel they are treated unfairly, they withhold their gifts or even actively retaliate against the firm (Akerlof, 1984; Yellen, 1984).

2.1 The evidence on gift exchange

Laboratory experiments support the gift exchange model (see Camerer and Weber (forthcoming) and references therein), and the proposition that past wage experiences can create wage reference points (Fehr et al., 2006). However, the relevance of some of these results to real employment is questionable: Gneezy and List (2006) show that positive reciprocity is fleeting in lengthy, real-effort tasks. Given the duration of most jobs, Gneezy and List argue that *positive* reciprocity cannot explain why newly hired workers are paid above-market wages, though it is still possible negative reciprocity

[2]See Katz (1986) for an overview of the different models.

is made of sterner stuff, and it is what keeps wages from falling.

This possible asymmetry between positive and negative reciprocity was explored in a recent field experiment by Kube et al. (2007). The researchers cut wages, or more accurately, frustrated wage expectations, by advertising the job using ambiguous language and then exploiting this ambiguity to pay some workers (college undergraduates) less than they expected. One limitation of this study is that the treatment depends upon confusion about the real wage, and this is not what wages cuts are really like. However, this treatment and a real wage cut are both examples of frustrated wage expectations, and we might reasonably expect that worker reactions will be similar in either case.

Although field experiments offer a more compelling test of gift exchange than the laboratory experiments, observational data from real, long-term employment scenarios would be more convincing, but the circumstances needed for causal inference—idiosyncratic factors leading to wage changes for one group of workers but not for another—are rare. Occasionally, this kind of scenario occurs, and the evidence from them is consistent with the view that negative reciprocity is long lasting and harmful to the organization. Mas (2006) found that New Jersey police forces losing arbitration closed fewer cases; Lee and Rupp (2007) found that airline pilots subject to large, industry-wide wage cuts were late more often. In the airline example, the effects were modest and transitory, perhaps because most airlines were at or near bankruptcy when the cuts were made, thus muting any fairness judgments.

2.2 Considering context

One common feature of the empirical evidence on wage cuts is that the context is either given by circumstance (the observational evidence and the field experiments) or the situation is decontextualized (the laboratory studies). Of course, every study has a particular context, and determining whether the findings are particular to the context is the inherent challenge of establishing external validity. But in the particular case of wage cuts, there are good reasons to think that context is especially important. Kahneman et al. (1986) found that people (a) view wage cuts that keep the firm solvent differently from cuts that increase already positive profits or to exploit market changes (b) see wage reference points as attached to specific workers in specific jobs, and not transferable to new employees or as having any relevance after sufficiently large reorganizations.

Context can be manipulated, either by design or by chance, but the only study

we are aware of exploiting variation in context is Greenberg (1990), who looked at employee theft in two factories following a temporary 15% pay cut. Employee theft rose in both plants, but in the plant where the CEO spent over an hour explaining the cuts and fielding questions, theft rose less than in the plant where management provided only a cursory explanation. Like Greenberg, we manipulate the context of the wage cuts, but we also look at a larger collection of outcomes, and because we conducted a controlled experiment, we have confidence that the response to the treatments is caused by framing of the wage cuts and is not an artifact of selection or the result of omitted variables.

2.3 A simple model of fairness judgments and quits

To discuss the experimental design and interpret the results, it is useful to consider a simple model of how factors that are economically irrelevant (i.e., do not affect payoffs) might still have economic consequences. Figure 1 illustrates a scenario where a worker is considering whether to continue at a piece-rate task, with the task having increasing marginal costs or whether to quit. Workers quit when the new wage offer is less than costs, with costs broadly defined to include the potential psychic cost of accepting an unfair wage.

Let c_0 and c_1 be the worker's cost for the initial task, additional tasks, respectively. In Figure 1, line b shows the worker's marginal benefit from working as a function of the wage. Ex ante, the worker's reservation wage for the initial task is w_0^r, corresponding to the point A where the benefit curve b intersects the cost curve c_0: $w_0^r = c_0$. The reservation wage for the additional task is w_1^r corresponding to the point B where the benefit curve intersects the cost curve c_1: $w_0^r = c_1$. Because $c_0 \leq c_1$, we can observe quits even without wage changes.

In a departure from the neoclassical model, we make three assumptions: (a) workers form reference points at previously received wages (b) workers find it psychologically painful to work for a wage below their reference wage and (c) this pain can be worsened or lessened by framing and context. If the employer parties the worker an above-market wage w_0^* initially, then the worker forms a reference point at E. "Unfair" wage offers increase the feeling of exploitation and thus the cost curve c_1^u has a steeper, negative slope than "fair" wage offers that create a flatter cost curve c_1^f. This difference in costs curves makes point C the reservation wage in the "fair" case and point D the reservation wage in the "unfair" case. In this example, the more unfair the wage offer, the more likely the worker is to reject it.

[Figure 1 about here.]

3 Experimental methodology

The experiment had three phases: (a) the wage expectation building phase (b) the treatment phase and the (c) the game phase. Figure 2 shows how subjects flow through the website hosting the experiment. In the wage expectation building phase, subjects transcribed three paragraphs, for 10 cents each. If a subject finished three paragraphs, they moved to the treatment phase. In the treatment phase, experiences differed: subjects in groups G2, G4, G6 and G7 were given the option to continue working, but at a lower wage, subjects in group G1 were given the same option, but at their previous wage while subjects in group G3 were not given the option to continue working. The framing of these various options also differed across groups. All subjects, regardless of what they did in the treatment phase, moved on to the games phase. In the games phase, they played a series of revealed-preference, contextualized games that measured their trust in us, patience and willingness to cooperate.

3.1 Experimental groups

The groups were:

G1 Control Subjects were offered their previous high wage of 10 cents to perform an additional task.

G2 Unexplained Subjects were offered a lower wage of 3 cents to perform an additional task. No explanation was given for the lower wage offer.

G3 No Cut Control Subjects were not offered an additional task. After the first set of tasks, they went straight to the follow-on games.

G4 Productivity Subjects were offered a lower wage of 3 cents to perform an additional task, but we first informed them that most workers can complete tasks more quickly after they gain some experience.

G5 Profit Subjects were offered a lower wage of 3 cents to perform an additional task, but we first informed them that we are trying to get as much work done for as little money as possible.

G6 Neighbor Subjects were offered a lower wage of 3 cents to perform an additional task, but we informed them that in our experience, other workers are willing to accept the lower offer.[3]

G7 Forced Subjects were *not* explicitly offered a lower wage to perform an additional task. Rather, they are simply brought to the next task, which was clearly labeled with the new, lower wage of 3 cents. There was a button at the bottom of the screen that allows them to quit and continue to the follow-on contextualized games.

G8 Primed Subjects were offered a lower wage of 3 cents to perform an additional task. No explanation was provided, but we first asked subjects what they thought would be a fair wage for doing an additional task.

Firms are likely to try justify their actions, which is why G4 is probably the most realistic treatment. No real firm would propose a wage cut in the same manner as the gratuitously insulting G5 treatment, yet it is still a useful treatment in that its effects might be similar to the effects of other actions real firms might take, such as proposing a wage cut following a large bonus for executives. In addition to hopefully angering workers, G5 tests whether any justification, not matter how insulting "works." G6 tests the notion that workers "learn" how to react from other workers. One reason why G6 might mitigate the behavioral response to a wage cut is that perhaps the psychic cost of taking work below one's reference point comes not just from a fear of being exploited, but also from a fear of being a singularly exploited. G7 tests whether not highlighting the wage cut and explicitly framing output for follow-on work as a binary decision affects behavior. G8 tests whether priming subjects to think about fairness enhances any effects.

[Figure 2 about here.]

3.2 Online marketplace as a testing domain

We recruited experimental subjects from an online labor market, created by a labor market intermediary (LMI) (Autor, 2008). Through an interface provided by the LMI, registered users perform tasks (posted by buyers) for money. The tasks are generally simple for humans to do yet difficult for computers—common tasks are captioning photographs, extracting data from scanned documents and transcribing audio clips.

[3]This was a true statement; there was no deception in the experiment.

Buyers control the features and contract terms of the tasks they post: they choose the design, piece-rate, time allowed per task, how long each task will be available and how many times they want a task completed.[4]

Workers, who are identified to buyers only by a unique string of letters and numbers, can inspect tasks and the offered terms before deciding whether to complete them. Buyers can require workers to have certain qualifications, but the default is that workers can "accept" a task immediately and begin work. Once the worker "submits" their work, buyers can approve or reject their submission. If the buyer approves, the LMI pays the worker with buyer-provided escrow funds; if the buyer rejects, the worker is paid nothing. The buyer can also grant bonuses, which is useful for our purposes since we can create complex contracts rather than use the default piece-rate format.

Although most buyers post tasks directly on the LMI website, it is possible to host tasks on an external site that workers reach by following a link. We used this external hosting method; we posted a single placeholder task containing a description of the work and a link to follow if subjects wanted to participate.

For the real-effort task, subjects transcribed (not translated) paragraph-sized chunks of Adam Smith's *The Wealth of Nations*. A sample paragraph is shown in Figure 3. This task was sufficiently tedious that no one was likely to do it "for fun" and it was sufficiently simple that all market participants could do the task. The source text was machine translated into Dutch, which increased the error rate of the transcriptions, thereby providing a more informative measure of work quality. Translating the text also prevented subjects from finding the text elsewhere on the Internet.[5]

[Figure 3 about here.]

3.3 Conditions for causal inference

To identify causal effects from the treatments, we needed ways to (a) as-good-as randomly assign subjects to different groups, (b) keep subjects from changing groups once assigned (c) keep subjects from participating multiple times, and (d) minimize non-random attrition.

[4]Tasks are often done multiple times by different workers for quality-control purposes.

[5]Since the text was presented as images, subjects were unable to simply copy and paste the text into the text box on the form. It is irrelevant whether or not any subjects actually knew Dutch since, if anything, knowledge of the language might make the task more difficult given the poor quality of the translation. One subject that apparently did speak Dutch sent an email warning us that the work was "grammatical gibberish."

For (a): We sequentially assigned subjects to treatment groups as they accepted our task. Unbeknownst to the subjects, clicking on the link associated them with a treatment group and redirected them to our site, thus stratifying subjects by arrival time. Because subjects were unaware of when other subjects arrived or the existence of other treatments, never mind the next treatment in the "queue," we are confident that treatment assignment was unconfounded with worker attributes.

For (b): If workers were aware of the different treatment groups, they would have an incentive to get into a group with a larger payoff. Even though subjects were unaware of the other treatments, to prevent the possibility of subjects hunting for the "best" treatment group, the software assigning subjects to groups tracked users' IP address. This tracking prevented subjects from breaking the randomization after their initial "assignment" click.

For (c): If a worker had multiple online identities, they could in principle complete our task multiple times. However, this double-dipping is unlikely: because buyers often want the same task done multiple times by different workers, the LMI designed several policies and software features to prevent a worker from having more than one account.[6]

For (d): To prevent differential attrition driven by differences among treatments, all subjects had identical initial experiences during the wage expectation building phase. Subjects' experiences differed by group only after already performing three transcriptions. Because of this investment, there was no attrition. An important point is that although we assigned subjects to treatments at the moment they accepted the task, the conceptual point of randomization was after the wage expectation stage but before the treatment stage.

The pre-randomization attrition is easy to spot in the data: attriting subjects just stop working before the end of the paragraph, never get the treatment stage and never submit their completion code to receive payment. We dropped from the sample any worker making more than 100 errors on a paragraph. This occurred for a little less than 20% of all subjects who started the paragraphs, but since these individuals were dropped prior to conceptual randomization, this kind of attrition is immaterial. Of the dropped subjects, none completed the survey, which gives us confidence that there were true attritors and not just bad performers.

[6]Workers must agree to have only one account—any detected attempt to have multiple accounts leads to a permanent ban. The LMI also requires browsers to accept cookies. If a person had two bank accounts or credit cards and two separate computers not sharing a network connection, they could in principle participate twice, but given the low stakes and risks involved (if this was detected, they would be banned from the site), we consider this possibility highly unlikely.

As a test of the randomization, we can see whether the distribution of collected covariates across the treatment groups is consistent with random assignment. For the binary covariates of gender, resident in the U.S. or Canada and whether the worker spends more than 10 hours a week online doing tasks for money, the Pearson's Chi-squared test p-values are 0.06, 0.12 and 0.21, respectively. While the low p-value for the gender is a little disconcerting, this is almost certainly the result of random chance and is not a major concern. However, since these measures are self-reported, there is a possibility that they are affected by the treatments, and for this reason all regressions are shown with and without the inclusion of controls.

3.4 Measuring outcomes following the wage cut

To measure quits, we observed how many subjects agreed to the follow-on wage offer; to measure work quality, we calculated how many edits would be required to transform the provided transcription provided transcription to a perfect transcription of the underlying text.[7] To measure patience, trust, and cooperation, we created contextualized games that that would not seem unusual to subjects.

To measure patience, we simply asked subjects whether they would be willing to forgo nearly immediate payment in exchange for a 30 cent bonus. We measured trust in the "firm" with a modified version of the trust game (Glaeser et al., 2000). To save money, we told subjects that we would select one player at random and implement his or her choices. Subjects were told to imagine they had $15 and that they could choose any fraction of that money to "send" to us. If we judged their work favorably, we would double what they sent and "send it back"—in the event of a unfavorable judgment, we would keep whatever they sent.[8] Although this game measures the worker's trust in us (e.g., trust that we will follow our own protocol and judge their work fairly), it is confounded with the worker's confidence, risk aversion and beliefs about our standards for the work. If these non-trust factors held constant across experimental groups we could in principle detect changes in trust, though this assumption of constant effects is questionable.

We measured cooperation in two ways: (a) willingness to give free advice and (b) willingness to take a short survey at a later date. If subjects agreed to take this follow-on survey, they were paid an additional 15 cents.[9] The survey measure

[7] This minimum-number-of-edits metric is called the edit distance, or the Levenshtein distance.

[8] The subject we randomly selected had chosen to send the full $15; we gave him or her $30.

[9] When you pay workers, you can embed a short message about why you are paying them—we embedded the link to our survey in this message. Remarkably, not one person who agreed to do the survey actually did it after we sent them the link. We had intended to use completing the survey as

of cooperation captures a mix of self-interest and cooperation with the firm, and is somewhat analogous analogous to a willingness to work overtime. For second cooperation measure, subjects were asked to suggest ways to make the transcription task easier. The wording of the request made it clear that offering advice was voluntary and that no payment would be made for the offered advice.[10]

4 Results

Table 1 reports the cross-tabulations for our main outcome measures (and binary covariates) by experimental group. Table 2 presents our results for quits and work quality, Table 3 trust and patience and Table 4 for cooperation. In all regressions, we used a linear model to analyze the data.[11] In Table 2, G2 is the excluded dummy variable—coefficients are marginal effects on the base rate in G2, "no explanation" case. We dropped data from G3 since these subjects were brought straight to the follow-on tasks. In Tables 3 and 4, no data was dropped and the "no offer" G3 group is excluded. We chose G3 because it approximates the status quo in the firm of no wage offers or changes.

For each of our post wage cut outcomes, we report regression results with and without the control variables. For the controls, we include an indicator for gender, whether the worker was from the U.S. or Canada, whether the worker self-reports spending more then 10 hours online per week working, whether the worker self-reports English as his or her first language and finally, the log of the sum of all the errors he or she makes on the first three paragraph transcriptions.

[Table 1 about here.]

4.1 Output—accepting the wage cut

Column Y of Table 2 reports estimated coefficients from an OLS regression of whether individuals accepted the wage cut on group assignment indicators; Column Y(c) re-

another measure of cooperation and trustworthiness, but besides there being no variation in survey response, we believe there may have been a technical error that prevented subjects from filling out the survey, given the complete lack of response.

[10]The text of the actual question was, "Please help us make these tasks easier for other workers. Describe what steps workers can take to make doing these transcriptions easier."

[11]Although some of our outcomes are dichotomous variables or counts, the linear model with robust standard errors offers several advantage over non-linear models: it is simpler, the coefficients are more easily interpretable and all treatment coefficients are identified, even when they perfectly predict success or failure (this is not the case with general linear models estimated by maximum-likelihood).

ports the results of the same regression when control variables are included. The regressions show that workers were more likely to do additional work when: we offered the previous wage (G1); we justified the wage cut in terms of increased productivity (G4); we indicated others are accepting the cut (G6); we failed to flag the wage cut as a new offer (G7). The G4 and G6 treatments reduced quits (i.e., increase the acceptance probability) by about .3, compared to when the cut is unexplained (G2). Uptake across G1, G4 and G6 was approximately the same, as can be seen by their associated regression coefficients or by Figure 4. For G7, the "forced" effect is about .6, which reflects the 100% uptake in G7 and the approximately 40% uptake in the comparison group.

Initially, we thought that G5 and G8 (the "profit" and "primed" treatments) would anger workers and cause more quits. While both treatments had a negative effect, they are not statistically significant. That G5 fails to reduce quits runs counter to studies showing that people acquiesce to demands when they are given a reason for that demand, no matter how inane the justification (for example, Langer et al. (1978) is the famous study showing that asking someone to let you cut in line to make copies "because you need to make copies" dramatically increases compliance).

The G7 "forced" group has higher output than the G1 group, even though the G7 offered wage was 70% lower. This surprised us, as we expected subjects to refuse the task because of the underhandedness of the offer. It is possible that subjects failed to consider their option of refusing to do the task, even though this option was clearly marked. Another possibility is a worker might believe they were in error: if a worker thought he had read the instructions incorrectly, he might worry that quitting would jeopardize payment for work already completed. This caveat aside, the evidence does suggest that explicitly flagging a wage change has a different effect than imposing the same change surreptitiously. This difference in reactions bolsters the notion that the higher salience of wage cuts compared to real wage erosion caused by inflation contributes to nominal wage rigidity.

[Figure 4 about here.]

[Table 2 about here.]

4.2 Quality—number of transcription errors

The columns in Table 2, where the headings contain E, report the OLS coefficients for a regression of logged error counts by group indicators and covariates. The only

strong result is that workers in the G5 "profit" group accepting the low wage offer made more errors; compared to the "no wage cut" excluded group, G5 subjects made, on average, 20 more errors. If we exclude an obvious outlier, the difference falls to about 8 errors. The outlier was one subject who agreed to the offer but left the text box blank. With and without the outlier, the effects are statistically significant ($p < .05$). While the finding looks like retaliation, it might also reflect selection, since the group have different quit rates and workers doing the fourth task might vary across treatments. Although we control for first-stage quality in Column E(c) and still find an effect, we are implicitly making a selection-on-observables assumption. Figure 5 is a scatter plot of the error count for every person. Although the G5 error points are higher, the selection issue is highlighted by the figure—there are far fewer data points.

We find no evidence that lower quality workers were less likely to quit; the coefficient on the logged errors is slightly negative in Column Y(c). This provides some evidence against the adverse selection model of efficiency wages (i.e., the best worker have the best outside options and are the most likely to quit after a wage cut). We also find no evidence that paying workers more improves quality; subjects in G1, the "no wage cut" group made slightly more errors, though the result is not statistically significant. This provides some evidence against the positive reciprocity conception of gift exchange.

[Figure 5 about here.]

4.3 Trust & patience—believing the firm, waiting for payment

Table 3, Column Trust reports the estimated coefficients from an OLS regression of the amount sent in a trust game by group indicators; Column Trust(c) includes controls. Generally, there does not appear to be any strong treatment effect on our trust measure. Subjects in the G6 "neighbor" group appear to trust more ($p < .05$), but this effect becomes smaller when estimated with controls, as well as less significant ($p < .10$). "Trust" is decreasing in first stage errors and is higher for men (though not at a statically significant level); both findings deepen our fear that our contextualized measure captures risk aversion and confidence.[12] Column Patient

[12]See Schubert et al. (1999) and references therein to gender differences and risk aversion. Although our measure was contextualized, it was framed in a way that made it seem more like a gamble than an investment, so finding a gender difference is consistent with both Schubert et al.'s experimental evidence and other papers on gender differences.

reports our coefficient estimates for the effects of the treatment on patience; we find no treatment effects, and only slight evidence that subjects from the US or Canada are less willing to accept a delay ($p < .10$).

4.4 Cooperation—offering advice

For the first cooperation measure, priming workers to think about fairness (G8) increases whether they give any advice, while mentioning that other workers are willing to take the wage cut (G6) seems to increase the amount of offered advice, though this evidence is not that strong. Table 4, Column Advice reports the results of a regression designed to measure effects on cooperation; the number of characters of advice offered by a subject is regressed on the treatment indicators. Column Advice (c) includes controls. Columns Advice 2 and Advice 2(c) correspond to the Column Advice and Column Advice (c) regressions, except the dependent variable is a binary indicator for whether the subject offered any advice at all.

For advice length, only G6, the "neighbor" treatment, appears to have an effect: in this group, subjects increased advice length by an average of about 50 characters ($p < .05$). The G6 "neighbor" treatment indicator is also significant when predicting whether any advice at all is offered ($p < .10$), but only when controls are included. In both the (Advice 2) and (Advice 2(c)) regressions, the G8 "priming" group indicator is significant ($p < .10$ without controls, $p < .05$ with controls), as is G2, "unexplained" is significant ($p < .10$ without controls; $p < .05$ with controls). Figure 6 plots the character length of the advice against the treatments. In this plot, the effect of G6 on the amount of advice can be seen in the cluster of points between 200 and 275, while the effects of G8 and G2 can be see in the relative absence of points at zero.

The G2 effect is surprising and we cannot think of a plausible hypothesis that explains this result, aside from random chance. For the G6 and G8 results, there is one speculative interpretation: by invoking social comparisons, G6 and G8 somehow put subjects in an accommodating mood. One common feature of both G6 and G8 is that they both explicitly invoke a social comparison (i.e., G6 introduces information about what other workers have done, while G7 discusses fairness). G4 discusses other workers' productivity, but this framing might be more about the nature of the task than about comparing oneself to other workers. Perhaps the G6 and G8 treatments somehow cause subjects to think more about other workers and therefore behave more charitably.

[Figure 6 about here.]

4.5 Cooperation—agreeing to a survey

For the second cooperation measure (willingness to take a follow-on survey), the "profit" treatment (G5) strongly reduces cooperation: only 74% of workers in G5 agree to take the survey, compared to an overall acceptance rate of 94% (not including G5). Table 4, Column Survey show that those G5 subjects have about a .25 lower probability of accepting the survey offer compared to the G3 comparison group ($p < .01$). What is interesting about this measure is that the presumably offended workers chose not to participate even though this would be a perfect way to retaliate—they could accept payment and then ignore the survey (since they would receive payment before receiving the survey link), yet they choose not to do this. Instead, they choose not to participate at all, foregoing an easy 15 cents.

[Table 3 about here.]

[Table 4 about here.]

5 Discussion

We find that the manner in which a wage cut offer is framed affects output, cooperation and possibly work quality. Our results undermine the notion that workers have a fixed reservation wage determined only by the onerousness of the work and the offered wage; workers' assessment of the fairness of the offer matters, and this assessment seems to depend upon contextual factors.

In terms of economic theory, our results are consistent with the negative reciprocity aspect of the gift exchange model. Our workers responded to wage cuts in ways damaging to the firm, though this damage was mediated by the framing of the cuts. If we had been a profit-maximizing firm, we might not have cut wages for the follow-on task, assuming our costs for transcription errors, turn-over and survey non-compliance were sufficiently high. Of course, the fact that firms do sometimes cut wages is proof that these worker-response considerations are not always paramount. One reason why the worker response might not override the benefits of cutting wages is that the situation is sufficiently dire that the firm will get a "pass" for cutting wages. We do find that seemingly valid justifications for a wage cut can help sweeten the bitter medicine, but workers are not fools, and offensive justifications actually make things worse. The main conclusion we draw from the evidence is that the notion that firms set wages in the shadow of the worker's taste for fairness has at least some explanatory power.

It is beyond the scope of this paper to analyze the welfare implications of worker fairness judgments and their effect on wage-setting, but we should note that the way fairness judgments seem to work do not, at least, exacerbate the business cycle: when firms must reduce costs, wage cuts are less likely to have the bad effects they would during booming times.

5.1 External validity

As with any study, one should be concerned about the external validity of our results. The differences between our setting and a real labor market are almost too numerous and obvious to mention.[13] Our setting does have the advantage over some laboratory studies in that subjects are performing real-effort tasks for money in what they believe is a real, albeit unusual, market setting. However, assessing external validity does not mean just tallying up the similarities and differences and reporting the "distance" between the experimental setting the real-world phenomena; assessing external validity requires one to consider how this "distance" is likely to alter the findings.

In our experimental setting, the short duration, anonymity and "one shot" nature of interactions should push subjects towards viewing everything through a homo economicus lens, and yet we find fairness judgments having strong effects. The incredibly low stakes cut has an ambiguous effect: on one hand, the wage cuts are absolutely small and thus perhaps less offensive, yet the small stakes make principled refusals less costly. In conventional jobs, quitting has far greater consequences and we would expect workers to be less sanguine about losing their jobs, even after a wage cut. The the higher stakes and longer duration of real jobs would probably make any fairness judgment more strongly felt. These more salient fairness judgments might magnify the response to wage cuts, though perhaps reactions with direct income consequences for the worker (such as quitting) are less likely and withdrawal of cooperation is more likely.

5.2 Interpreting employment as an incomplete contract

Although we find some limited evidence of low work quality coinciding with the apparently offensive "profit" justification, our holistic interpretation of the results is

[13] One distinction between our scenario and real scenarios is that our tasks are piece-rate, while most real employers pay hourly wages. While we think this distinction is not relevant to the whether the cut triggers a fairness judgment, piece-rate work does raise some strategic issues (Carmichael and MacLeod, 2000).

that "retaliation" or "negative reciprocity" are probably the wrong characterizations. Although the term "negative reciprocity" conjures images of employees pilfering supplies or sabotaging equipment, Bewley (1999) found that what firms fear is not overt retaliation; they fear that morale will sag and that disgruntled workers will no longer identify with with the firm or adopt its goals as their own. The incomplete contracting model introduced by Hart and Moore (2008) offers a better conception of what sours in the employment relationship after a wage cut. In the Hart-Moore model, the performing party makes a non-contractible decision as to whether she will provide consummate performance or to-the-letter, perfunctory performance. For the performing party, the added costs of consummate performance are negligible but can have large implications for the buyer.[14]

The inability of a buyer to compel consummate performance is an example of the broader principal-agent problem and is common in modern models of employment. To connect this feature back to Bewley's survey research, it is interesting that many managers were worried that after a wage cut, employees would no longer "identify" with the firm. To "identify" with someone in a psychological sense is to have empathy for that person and to see your interests and concerns as coinciding. This alignment of interests between firm and employee, even if more psychological than financial, seems like a way to overcome the inherent principal-agent problem of employment. Losing this solution to the principal-agent problem is one way of conceptualizing precisely what firms fear will happen if they cut wages.

6 Conclusion

Our main positive finding is that framing can affect the behavioral response to a wage change. These worker reactions are fairly easy to understand: workers reduce output and cooperation when wages are cut for capricious or selfish-seeming reasons, but workers do not reduce output as much and apparently are still willing to cooperate if the employer has seemingly valid reasons for his or her actions.

Although this paper focuses on the role of fairness judgments in causing nominal wage rigidity, it is not the only example of moral judgments shaping the contours of the economy. It seems likely that popular moral sentiment and human psychology shape laws, institutions and customs. Roth (2007), which looks at repugnance as a

[14]For example, consider a worker noticing that an expensive piece of machinery that she is not directly responsible for is about to run out of oil and become damaged; alerting someone takes almost no effort for the employee but could save the firm a great deal of money.

constraint on markets, is one example of the larger phenomena of what might be called positive moral philosophy affecting economic life. More research in this vein might lead to a better understanding of the "folk morality" people have about markets. It would be particularly interesting, and perhaps useful to policy-makers, to know what kinds of moral judgments are universal and seemingly invariant and which are mutable.

References

Akerlof, George A., "Gift Exchange and Efficiency-Wage Theory: Four Views," *The American Economic Review*, 1984, *74* (2), 79–83.

Autor, David H., "The Economics of Labor Market Intermediation: An Analytic Framework," Working Paper 14348, National Bureau of Economic Research September 2008.

Camerer, Colin and Roberto Weber, "Experimental Organizational Economics," *Handbook of Experimental Economics*, forthcoming.

Carmichael, H.L. and W.B. MacLeod, "Worker cooperation and the ratchet effect," *Journal of Labor Economics*, 2000, *18* (1), 1–19.

Fehr, Ernst, Armin Falk, and Christian Zehnder, "Fairness Perceptions and Reservation Wages — The Behavioral Effects of Minimum Wage Laws," *Quarterly Journal of Economics*, 2006, *121* (4).

Glaeser, Edward L., David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter, "Measuring Trust," *The Quarterly Journal of Economics*, 2000, *115* (3), 811–846.

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.

Greenberg, J., "Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts," *Journal of Applied Psychology*, 1990, *75* (5), 561–568.

Hart, Oliver and John Moore, "Contracts as Reference Points," *Quarterly Journal of Economics*, 2008, *123* (1), 1–48.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *The American Economic Review*, 1986, *76* (4), 728–741.

Katz, Lawrence F., "Efficiency Wage Theories: A Partial Evaluation," NBER Working Papers 1906, National Bureau of Economic Research, Inc April 1986.

Kube, Sebastian, Michel A. Maréchal, and Clemens Puppe, "Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment," *Working Paper*, 2007.

Langer, Ellen, Arthur Blank, and Benzion Chanowitz, "The Mindlessness of Ostensibly Thoughtful Action: The Role of "Placebic" Information in Interpersonal Interaction.," *Journal of Personality and Social Psychology*, 1978, *36* (6), 635–42.

Lee, Darin and Nathanial G. Rupp, "Retracting a Gift: How Does Employee Effort Respond to Wage Reductions?," *Journal of Labor Economics*, 2007, *25* (4), 725–761.

Mas, Alexandre, "Pay, Reference Points, and Police Performance," *The Quarterly Journal of Economics*, 2006, *121* (3), 783–821.

Roth, Alvin E., "Repugnance as a Constraint on Markets," *Journal of Economic Perspectives*, 2007, *21* (3), 37–58.

Schubert, Renate, Martin Brown, Matthias Gysler, and Hans Wolfgang Brachinger, "Financial Decision-Making: Are Women Really More Risk-Averse?," *The American Economic Review*, 1999, *89* (2), 381–385.

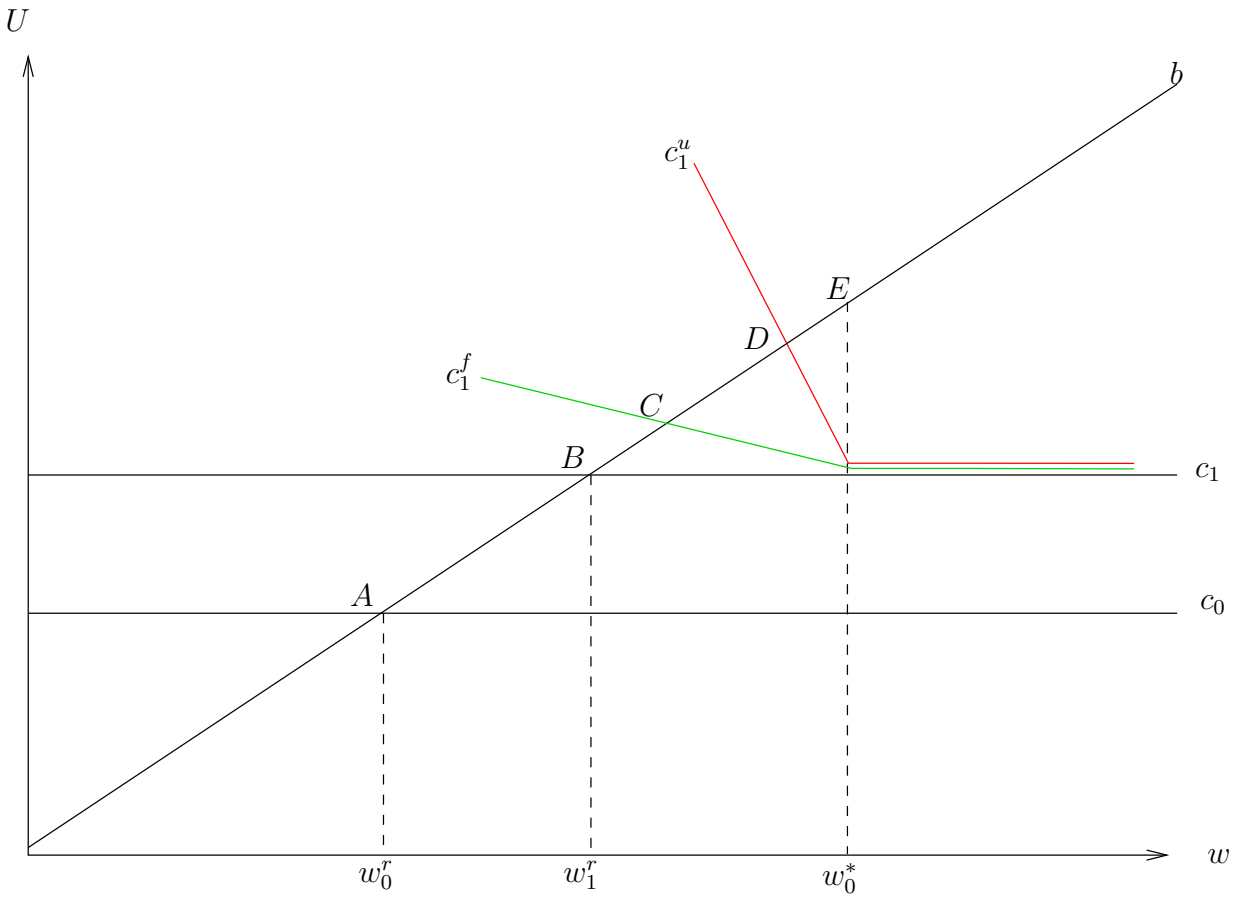
Shapiro, Carl and Joseph E. Stiglitz, "Equilibrium Unemployment as a Worker Discipline Device," *The American Economic Review*, 1984, *74* (3), 433–444.

Truman, F. Bewley, *Why Wages Don't Fall During a Recession*, Harvard University Press, 1999.

Yellen, Janet L., "Efficiency Wage Models of Unemployment," *American Economic Review*, 1984, *74* (2), 200–205.

List of Figures

Figure 1: Quits following new wage offers after worker has altered reference point due to past wage experience.



Notes: This figure illustrates how being paid above one's reservation wage initially (a) can change the reservation wage for subsequent offers, depending upon how that offer is framed. If the task has increasing marginal costs, then the initial per-task performance cost is c_0, which is below c_1, the costs for the follow-on task. We assume a constant marginal utility of wealth, so from the 1st task to the second task, the neoclassical reservation wage moves from A to B. However, if the worker is paid w_0^* for the initial task, the reference point created by this experience might affect the reservation wage for subsequent offers. For an unfair offer (in red), the performance cost increases to c_1^u and the reservation wage is D. With a fairer offer (in green), the cost curve is c_1^f and reservation wage is at C, which is still above the neo-classical reservation wage at B for this task, but lower than the the "offensive" offer reservation wage at D.

Figure 2: Experimental Design

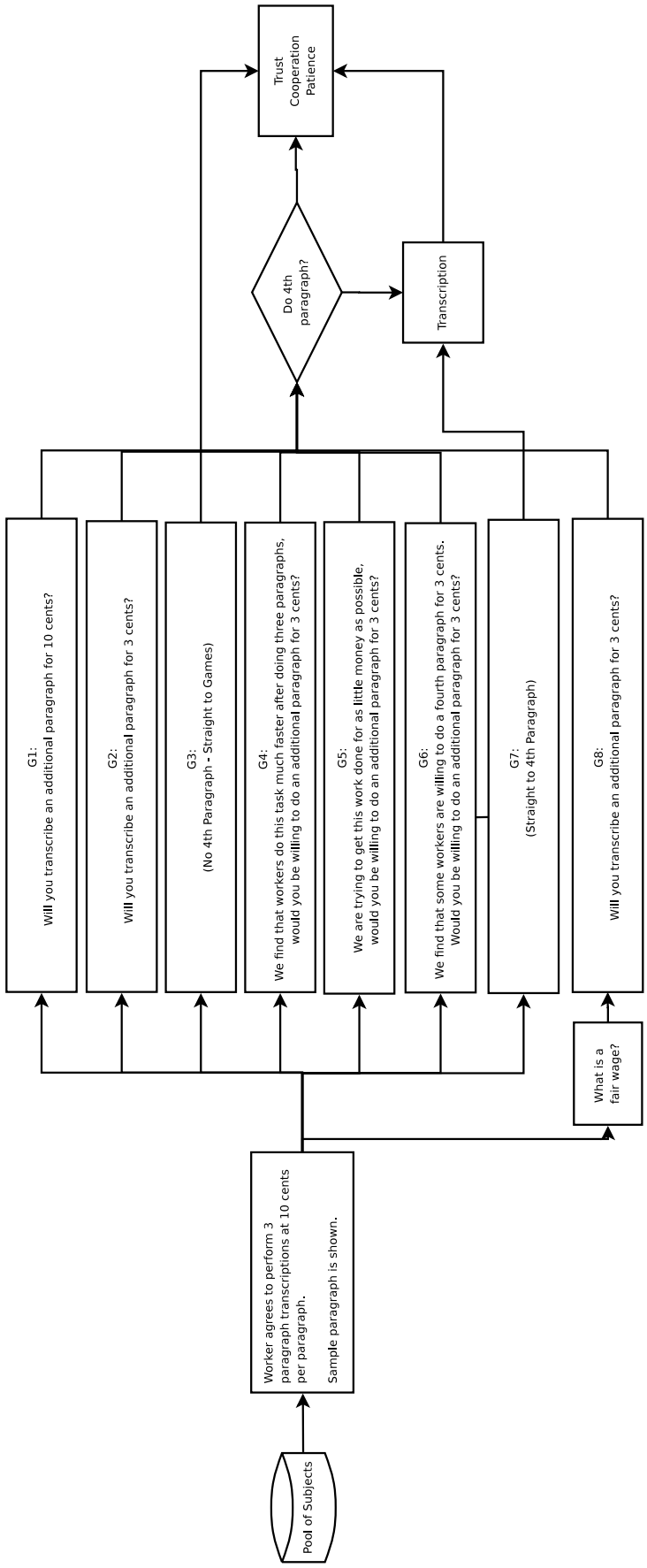
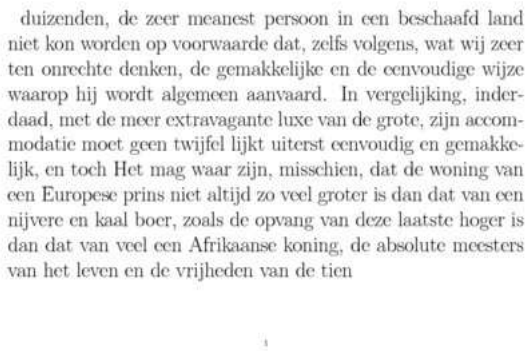
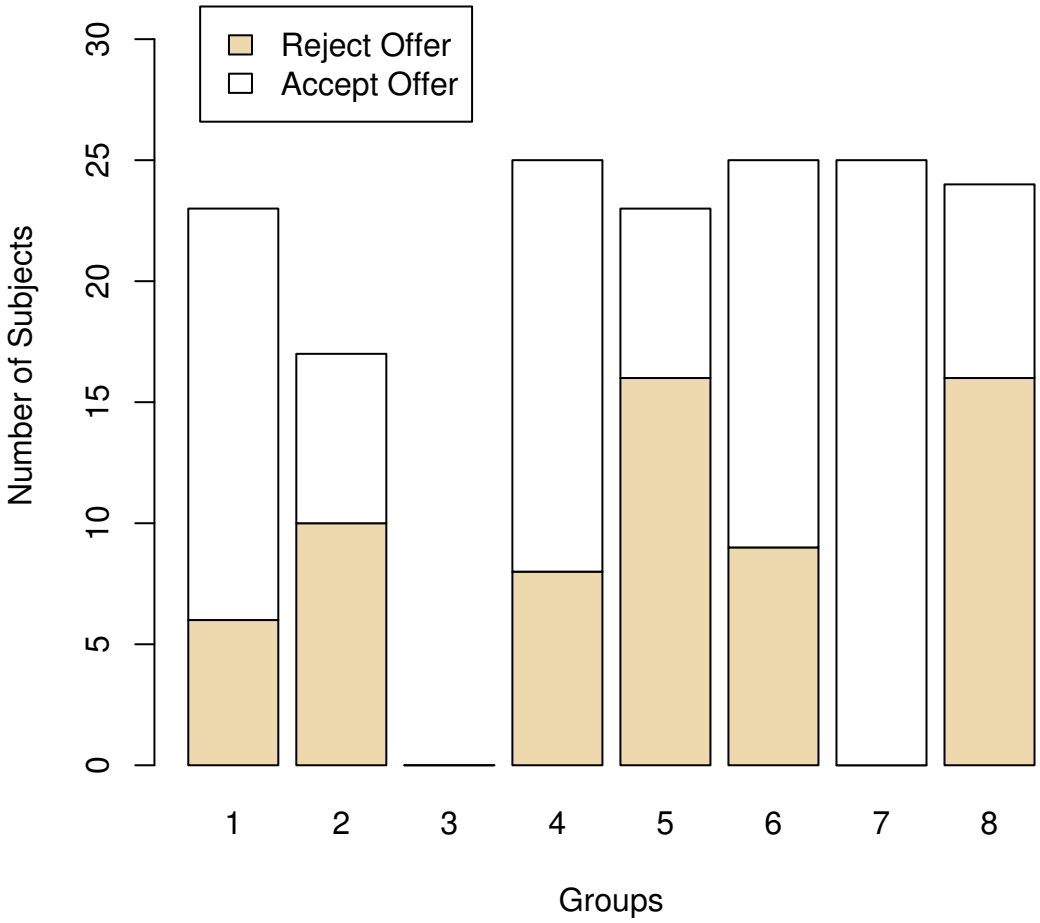


Figure 3: Sample Real-effort, piece-rate task


duizenden, de zeer meanest persoon in een beschaafd land niet kon worden op voorwaarde dat, zelfs volgens, wat wij zeer ten onrechte denken, de gemakkelijke en de eenvoudige wijze waarop hij wordt algemeen aanvaard. In vergelijking, inderdaad, met de meer extravagante luxe van de grote, zijn accommodatie moet geen twijfel lijkt uiterst eenvoudig en gemakkelijk, en toch Het mag waar zijn, misschien, dat de woning van een Europese prins niet altijd zo veel groter is dan dat van een nijvere en kaal boer, zoals de opvang van deze laatste hoger is dan dat van veel een Afrikaanse koning, de absolute meesters van het leven en de vrijheden van de tien

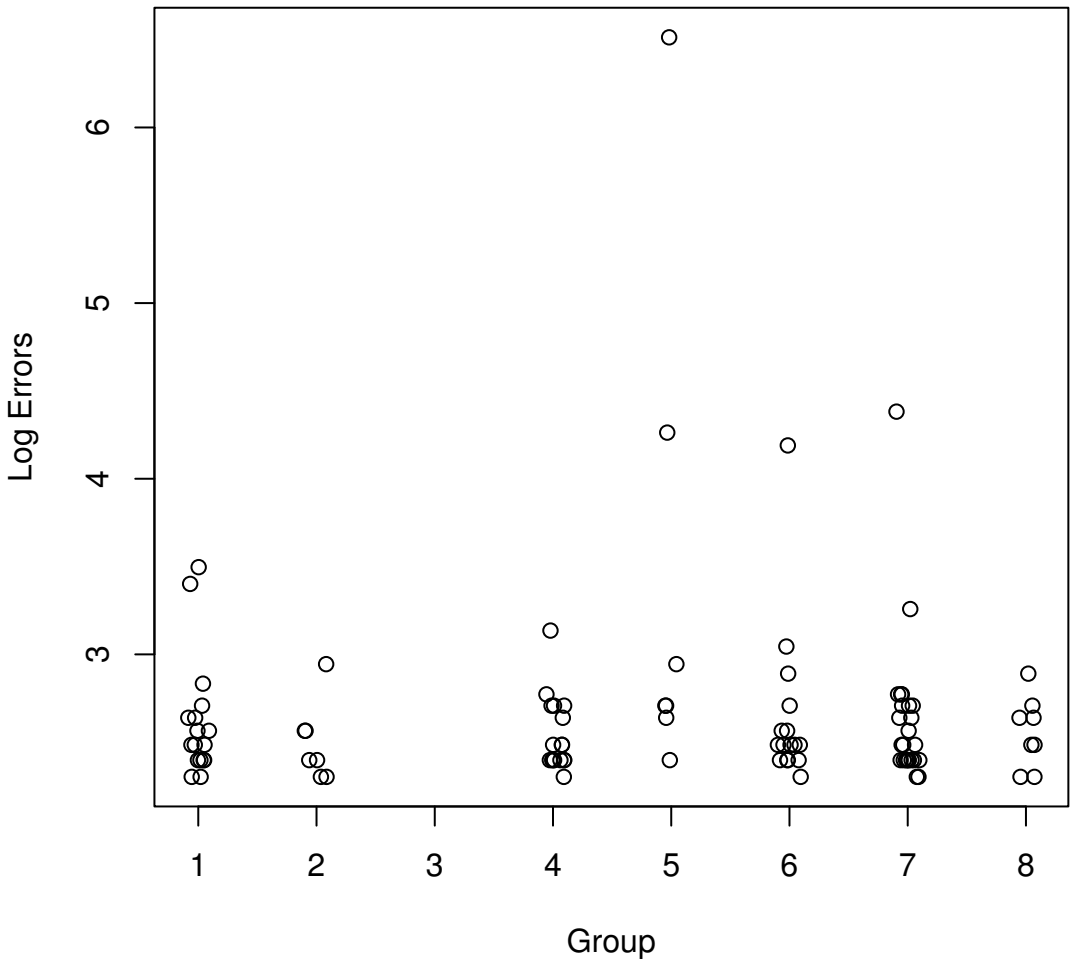
Notes: Subjects were asked to transcribe paragraphs like the one above into a text box. It is a short selection from the "Wealth of Nations" that we machine-translated into Dutch.

Figure 4: Quits after first stage by group



Notes: G1 Control—Offered same rate (not a wage cut)
G2 Unexplained—No explanation for low offer
G4 Productivity —Low offer explained as response to increased productivity
G5 Profit—Low offer explained as profit motive by employer
G6 Neighbor—Subjects told others willing to accept the offer
G7 Forced—Subjects brought to next task without receiving an offer
G8 Primed for Fairness—Subjects asked what is a fair wage before offer

Figure 5: Log errors on fourth paragraph, with errors defined as the minimum edit distance



Notes: The horizontal coordinate of the data points are "jiggered" (i.e., randomly perturbed), to reveal point density.

G1 Control—Offered same rate (not a wage cut)

G2 Unexplained—No explanation for low offer

G4 Productivity—Low offer explained as response to increased productivity

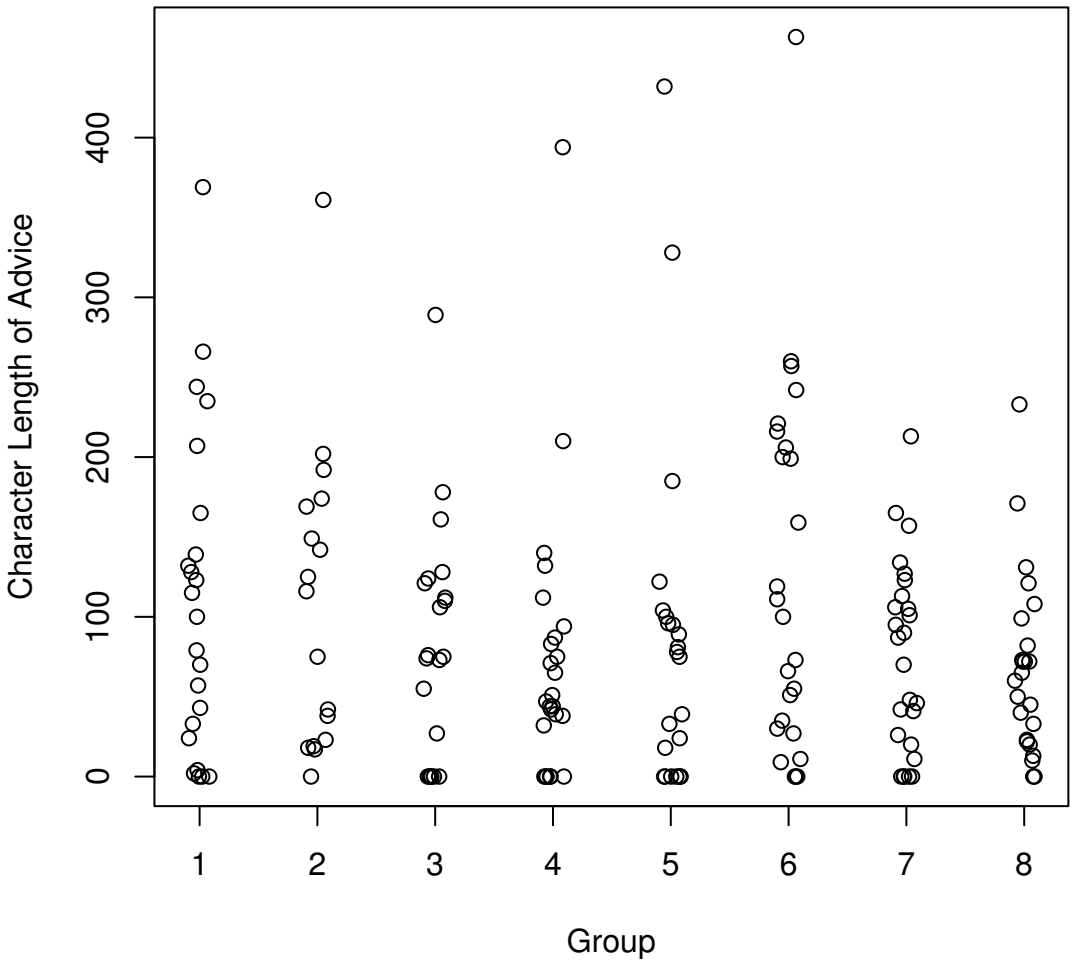
G5 Profit—Low offer explained as profit motive by employer

G6 Neighbor—Subjects told others willing to accept the offer

G7 Forced—Subjects brought to next task without receiving an offer

G8 Primed for Fairness—Subjects asked what is a fair wage before offer

Figure 6: Advice length, in number of characters, by group



Notes: The horizontal coordinate of the data points are "jiggered" (i.e., randomly perturbed), to reveal point density.

G1 Control — Offered same rate (not a wage cut)

G2 Unexplained—No explanation for low offer

G3 No Cut Control—Subjects go straight to games

G4 Productivity—Low offer explained as response to increased productivity

G5 Profit—Low offer explained as profit motive by employer

G6 Neighbor—Subjects told others willing to accept the offer

G7 Forced—Subjects brought to next task without receiving an offer

G8 Primed for Fairness — Subjects asked what is a fair wage before offer

List of Tables

Table 1: Cross Tabulations by Group

	N	1	2	3	4	5	6	7	8
		$N = 23$	$N = 17$	$N = 21$	$N = 25$	$N = 23$	$N = 25$	$N = 26$	$N = 24$
U.S./Can.	184	61% (14)	94% (16)	86% (18)	76% (19)	70% (16)	64% (16)	85% (22)	83% (20)
Male	184	35% (8)	12% (2)	33% (7)	28% (7)	48% (11)	40% (10)	38% (10)	8% (2)
$>$ 10 hrs.	184	39% (9)	41% (7)	62% (13)	64% (16)	43% (10)	36% (9)	31% (8)	50% (12)
Accepts wage offer	162	74% (17)	41% (7)		68% (17)	30% (7)	64% (16)	100% (25)	33% (8)
Co-operates (agrees to survey)	184	91% (21)	94% (16)	100% (21)	84% (21)	74% (17)	100% (25)	96% (25)	92% (22)
Patient	184	61% (14)	65% (11)	71% (15)	72% (18)	65% (15)	84% (21)	81% (21)	79% (19)
Co-operates (offers advice)	184	87% (20)	94% (16)	71% (15)	76% (19)	70% (16)	88% (22)	81% (21)	92% (22)
Trusts (sends $>$ 7.5)	184	48% (11)	53% (9)	38% (8)	40% (10)	57% (13)	68% (17)	42% (11)	42% (10)

Notes: N is the number of non-missing observations and the number after the percents are frequencies. *add* is an indicator for whether the worker performed an additional task after the pay cut offer.

G1 Control—Offered same rate (not a wage cut)

G2 Unexplained—No explanation for low offer

G3 No Cut Control—Subjects go straight to games

G4 Productivity —Low offer explained as response to increased productivity

G5 Profit—Low offer explained as profit motive by employer

G6 Neighbor—Subjects told others willing to accept the offer

G7 Forced—Subjects brought to next task without receiving an offer

G8 Primed for Fairness—Subjects asked what is a fair wage before offer

Table 2: Estimated coefficients from an OLS regression of output (Y) and log errors (E) on experimental group indicators and subject characteristics

	Y	Y(c)	E	E(c,e)	E(c)	E(no out)
Intercept	0.41***	0.73*	2.50***	2.36***	0.85†	1.30***
	(0.11)	(0.31)	(0.19)	(0.26)	(0.50)	(0.36)
G1	0.33*	0.40**	0.12	0.15	0.17	0.11
	(0.14)	(0.14)	(0.23)	(0.24)	(0.23)	(0.16)
G4	0.27†	0.31*	0.05	0.08	0.00	−0.00
	(0.14)	(0.14)	(0.23)	(0.24)	(0.22)	(0.16)
G5	−0.11	−0.01	0.96***	0.96**	1.01***	0.48*
	(0.14)	(0.14)	(0.27)	(0.28)	(0.27)	(0.20)
G6	0.23	0.31*	0.15	0.16	0.05	0.03
	(0.14)	(0.14)	(0.23)	(0.24)	(0.23)	(0.17)
G7	0.59***	0.67***	0.13	0.11	0.05	0.04
	(0.14)	(0.14)	(0.22)	(0.23)	(0.22)	(0.15)
G8	−0.08	−0.08	0.06	0.13	0.07	0.04
	(0.14)	(0.14)	(0.27)	(0.28)	(0.26)	(0.19)
Male		−0.18*		0.12	0.12	−0.02
		(0.08)		(0.13)	(0.12)	(0.09)
U.S./Can.		0.15		0.03	0.11	−0.00
		(0.11)		(0.17)	(0.16)	(0.12)
$>$ 10 hrs.		0.08		−0.03	−0.04	−0.15†
		(0.07)		(0.12)	(0.11)	(0.08)
English		−0.18†		0.11	0.05	0.03
		(0.11)		(0.17)	(0.16)	(0.12)
Log Errors		−0.08			0.40***	0.34***
		(0.07)			(0.11)	(0.08)
N	162	162	97	97	97	96
R^2	0.23	0.28	0.17	0.19	0.29	0.25
adj. R^2	0.20	0.23	0.12	0.10	0.20	0.15

Standard errors in parentheses

Notes: Robust standard errors are in parentheses.

Significance Symbols: $\dagger : p \leq .10$, $* : p \leq .05$, $** : p \leq .01$ $* * * : p \leq .001$.

The excluded treatment is G2.

G1 Control — Offered same rate (not a wage cut)

G2 Unexplained - No explanation for low offer

G4 Productivity — Low offer explained as response to increased productivity

G5 Profit — Low offer explained as profit motive by employer

G6 Neighbor — Subjects told others willing to accept the offer

G7 Forced — Subjects brought to next task without receiving an offer

G8 Primed for Fairness — Subjects asked what is a fair wage

Table 3: Estimated coefficients from an OLS regression of trust and patience measures by group indicators and subject characteristics.

	Trust	Trust(c)	Patient	Patient(c)
Intercept	5.81***	11.91***	0.71***	1.00**
	(1.06)	(3.24)	(0.10)	(0.30)
G1	1.71	1.08	−0.11	−0.17
	(1.47)	(1.50)	(0.14)	(0.14)
G2	2.48	2.51	−0.07	−0.10
	(1.59)	(1.59)	(0.15)	(0.15)
G4	1.19	1.23	0.01	−0.03
	(1.44)	(1.44)	(0.13)	(0.13)
G5	2.67†	2.25	−0.06	−0.09
	(1.47)	(1.49)	(0.14)	(0.14)
G6	3.15*	2.75†	0.13	0.09
	(1.44)	(1.47)	(0.13)	(0.14)
G7	2.11	1.81	0.09	0.07
	(1.43)	(1.46)	(0.13)	(0.13)
G8	1.57	1.69	0.08	0.04
	(1.45)	(1.47)	(0.13)	(0.14)
Male		0.53		−0.12
		(0.85)		(0.08)
U.S./Can.		−1.35		−0.20†
		(1.15)		(0.11)
$>$ 10 hrs.		−1.13		−0.04
		(0.75)		(0.07)
English		0.37		0.14
		(1.11)		(0.10)
Log Errors		−1.24		−0.04
		(0.75)		(0.07)
N	184	184	184	184
R^2	0.04	0.08	0.03	0.06
adj. R^2	−0.00	0.01	−0.01	−0.00

Notes: Robust standard errors are in parentheses. Significance Symbols: $\dagger : p \leq .10$, $* : p \leq .05$, $** : p \leq .01$ $*** : p \leq .001$.
G3 is the omitted factor in all regressions.
G1 Control—Offered same rate (not a wage cut)
G2 Unexplained—No explanation for low offer
G3 No Cut Control—Subjects go straight to games (excluded)
G4 Productivity—Low offer explained as response to increased productivity
G5 Profit—Low offer explained as profit motive by employer
G6 Neighbor—Subjects told others willing to accept the offer
G7 Forced—Subjects brought to next task without receiving an offer
G8 Primed for Fairness—Subjects asked what is a fair wage

Table 4: Estimated coefficients from an OLS regression of cooperation measures by group indicators and subject characteristics.

	Advice	Advice(c)	Advice 2	Advice 2(c)	Survey	Survey(c)
Intercept	81.38***	132.31*	0.71***	0.78**	1.00***	1.27***
	(19.34)	(58.35)	(0.08)	(0.26)	(0.06)	(0.18)
G1	28.84	29.80	0.16	0.18	−0.09	−0.06
	(26.75)	(27.11)	(0.12)	(0.12)	(0.08)	(0.08)
G2	28.15	35.06	0.23†	0.26*	−0.06	−0.04
	(28.91)	(28.75)	(0.12)	(0.13)	(0.09)	(0.09)
G4	−9.38	−8.99	0.05	0.06	−0.16†	−0.13
	(26.23)	(26.04)	(0.11)	(0.11)	(0.08)	(0.08)
G5	1.18	−0.72	−0.02	−0.01	−0.26**	−0.24**
	(26.75)	(26.86)	(0.12)	(0.12)	(0.08)	(0.08)
G6	43.02	50.20†	0.17	0.20†	−0.00	0.04
	(26.23)	(26.50)	(0.11)	(0.12)	(0.08)	(0.08)
G7	−7.53	−7.22	0.09	0.11	−0.04	−0.01
	(26.00)	(26.25)	(0.11)	(0.11)	(0.08)	(0.08)
G8	−14.09	−3.37	0.20†	0.25*	−0.08	−0.05
	(26.48)	(26.46)	(0.11)	(0.12)	(0.08)	(0.08)
Male		26.88†		0.13†		0.06
		(15.29)		(0.07)		(0.05)
U.S./Can.		−4.37		0.06		0.10
		(20.74)		(0.09)		(0.06)
> 10 hrs.		17.16		0.08		0.04
		(13.50)		(0.06)		(0.04)
English		38.94†		0.05		−0.03
		(19.96)		(0.09)		(0.06)
Log Errors		−24.91†		−0.06		−0.10*
		(13.59)		(0.06)		(0.04)
N	184	184	184	184	184	184
R^2	0.05	0.11	0.05	0.08	0.09	0.13
adj. R^2	0.01	0.05	0.01	0.02	0.05	0.07

Standard errors in parentheses

Notes: Robust standard errors are in parentheses. Significance Symbols: † : $p \leq .10$, ∗ : $p \leq .05$, ∗∗ : $p \leq .01$ ∗ ∗ ∗ : $p \leq .001$.

G3 is the omitted factor in all regressions.

G1 Control—Offered same rate (not a wage cut)

G2 Unexplained—No explanation for low offer

G3 No Cut Control—Subjects go straight to games (excluded)

G4 Productivity—Low offer explained as response to increased productivity

G5 Profit—Low offer explained as profit motive by employer

G6 Neighbor—Subjects told others willing to accept the offer

G7 Forced—Subjects brought to next task without receiving an offer

G8 Primed for Fairness—Subjects asked what is a fair wage

Procurement, Incentives and Bargaining Friction: Evidence from Government Contracts

John Horton*

Abstract

A transaction cost theory of procurement developed by Bajari and Tadelis (2001) models the buyer's choice of incentive structure as endogenous, with buyers trading off the efficiency of high-powered incentives against the ex post bargaining friction these incentives can create. The source of the bargaining friction is assumed to be asymmetric information between the buyer and seller about the true costs of adapting the project to changed conditions. Using government contract data and an instrument based on contracting-office idiosyncratic variation in preference for various contractual forms, I estimate the effect of a buyer choosing a fixed-price (i.e., high-powered) contract on the probability that the contract will lead to litigation, which proxies for bargaining failure and friction. I find that (a) fixed-price contracts are far more likely to be litigated that cost-plus contracts and (b) the instrumental variables estimate of the effect of choosing a fixed-price structure is almost twice as larges as the biased, OLS estimate. These results are consistent with the main predictions of the Bajari and Tadelis model.

1 Introduction

The key insight from transaction cost economics (TCE) is that the governance structure that controls the relationship between a buyer and a seller can create incentives for parties to engage in surplus-dissipating ex post strategic behavior. When the parties have some control over that governance structure, they will rationally take steps ex ante to mitigate these costs, such as by vertically integrating or writing long-term contracts. Despite a large body of empirical evidence suggesting the relevance of transaction costs, the informality of most TCE models makes discriminating tests of the theory difficult (Whinston, 2001). For example, there are several competing theories of the firm, and not surprisingly, all make the prediction of vertical integration. Given the limited set of

*John F. Kennedy School of Government, Harvard University

observable characteristics normally available to the researcher, it is usually hard to rule out alternate theories as the cause of some governance structure, never mind determine (a) whether transactions costs are truly determining that structure or (b) which precise transaction costs that are most important. A novel contribution of this paper is the use an actual transaction cost — namely litigation — as the outcome of interest rather than a second-order effect of a transaction cost such as vertical integration. Facilitated by a formal theory that makes precise predictions, this innovation allows a much stronger test than past empirical investigations of TCE theories.

1.1 The Choice of Contract Structure

Although contract incentive schemes can be arbitrarily complex in principle, most real-world contracts are simple, with fixed-price or cost-plus being the dominant contractual forms (Eggleston et al., 2000). A fixed-price contract requires the seller to agree to perform some specific task in exchange for a predetermined lump-sum payment, while a cost-plus contract requires the seller to agree to an (approximately) undefined amount of work in exchange for reimbursement of costs plus a fee. In general, fixed-price contracts provide sellers with greater incentive to reduce costs than cost-plus contracts do, but they tend to be less flexible and more contentious. Table 1 highlights the key qualitative differences between cost-plus and fixed-price contracts.

The simplicity of real contracts is at odds with the thrust of modern procurement theory, as exemplified by Laffont and Tirole (1993). In these theories' models, the seller's private, ex ante information about production costs is assumed to be the primary obstacle to creating successful contractual relationships. The remedy to this asymmetry is a menu of complex contracts designed to screen sellers by type. Though these models are intended to be normative, even if buyers did offer menus of contracts (which, generally speaking, they do not), the limited repertoire of fixed-price and cost-plus contracts seems insufficient for the delicate task of screening sellers. Bajari and Tadelis and others have suggested that even if ex ante asymmetries do exist, the mechanism for addressing the problem is not the choice of incentive structure, but rather some combination of reputation, competitive bidding, and the use of sureties (Bajari and Tadelis, 2001; Banerjee and Duflo, 2000).

An alternative transaction cost procurement model developed by Bajari and Tadelis is more consistent with the stylized facts of contracting (Bajari and Tadelis, 2001). Their model assumes that buyers and sellers are restricted to choosing simple linear contracts and that both parties are equally ignorant ex ante about production costs, but that as work on a task progresses, the seller has better information about the costs of the contingencies that inevitably arise. With a cost-

plus contract, contingencies — i.e., states not considered by the parties beforehand — pose no real problem because the contract already specifies the correct action, namely, that the seller receives his or her costs. In contrast, with a fixed-price contract, unforeseen[1] contingencies force the buyer and seller to reach a new agreement about cost, but the seller's private information creates bargaining friction and ultimately litigation if terms cannot be reached. In their model, the fundamental problem of contracting is not ex ante information asymmetry, but ex post asymmetry about the costs of adaptation and the resulting bargaining friction.

One of the many merits of the Bajari and Tadelis model is that its formalism generates several clear, unambiguous predictions. Two predictions that are the focus on this paper are: (a) fixed-price contracts should lead more often to bargaining failure than cost-plus contracts do, and (b) fixed-price contracts will be used more frequently when the good or service is relatively simple and when the buyer's expectations are easy to specify. The first prediction is a direct implication of the ex post bargaining requirement of the fixed-price contract, which simply produces a greater flow of potentially litigated disputes. To understand the second prediction, note that a fixed-price contract induces the greatest effort by the seller to control costs, but to minimize back-end bargaining friction costs inherent with the fixed-price structure, more resources must be spent by the buyer up front to enumerate contingencies and write a careful contract. Goods and services for which the buyer can easily specify his or her expectations and likely contingencies will have lower front-end contract writing costs and hence, all else being equal, are the kinds of goods and services for which the fixed-price contract is most attractive.

This paper tests the Bajari and Tadelis model by estimating of the effect of contract structure on the probability of bargaining failure using government procurement data, with litigation between the government and the firm as a proxy for bargaining friction. Because the choice of structure is likely to be endogenous even without a theory whose key prediction is the endogeneity of incentive structure, to identify the effect I exploit idiosyncratic variation in how different government contracting offices procure identical goods. I find that the use of a fixed-price contract strongly increases the probability of bargaining failure as measured by contract litigation. I also find that the unbiased instrumental variables estimate of the effect is almost twice as large as the ordinary least-squares (OLS) estimate. The downward bias of the OLS estimate is consistent with the notion that fixed-price contracts are used to procure goods or services that can be easily specified and therefore simple and inherently less dispute-prone. These two findings are consistent with the main predictions of the Bajari and

[1] Unforeseen in the sense that the buyer and seller did not write contract actions for this contingency — this is not a bounded rationality assumption.

Tadelis model, though they do not exclude other possible causal mechanisms, which are discussed in Section 5.1.

2 Theory

2.1 Litigation as Bargaining Failure

Because litigation is a costly, negative-sum game, rational parties with symmetric information would presumably never litigate disputes. Instead, potential litigants could always reach some negotiated settlement that, in expectation, would make both parties better off (Spier, 2008). Litigation models explain this "paradox" of litigation by assuming some kind of asymmetry between the parties, either informational, such as in their respective estimates of the probability of a trial outcome, or in their valuation of outcomes.[2] There is an extensive theoretical and empirical literature on how disputes are selected for litigation, with some substantive disagreements about assumptions and implications, but the key insight of this literature relevant here is that litigated cases are not just random samples from all cases; rather, litigation comes from a failure to reach a settlement, and this failure can depend strongly on the characteristics of the situation (Priest and Klein, 1984; Bebchuk, 1984; Eisenberg, 1990; Hylton, 1993; Kessler et al., 1996; Shavell, 1996). In this example, the incentive structure is hypothesized to affect ex post bargaining, and to lead to disputes.

2.2 The Bajari and Tadelis Model

In the Bajari and Tadelis model, buyers are restricted to offering simple, linear incentive contracts of the form $\alpha + \beta c$, where c is the seller's actual total costs, which are assumed to be verifiable ex post in total but not in part.[3] If $\beta = 0$, the contract is a fixed-price contract, while $\beta \geq 1$ is a cost-plus contract.[4] The buyer has a project that gives him utility v if it is completed and 0 otherwise. The assumption on which the entire model hinges is that in any project, there are multiple contingencies that might arise once work begins that will affect the seller's realized cost, and these realized costs are the seller's *private* information. Examples of such states include design or regulatory changes,

[2]For example, a company facing a discrimination lawsuit might have a high incentive to settle all but the weakest cases because the reputational costs of losing at trial might greatly exceed the more modest stakes perceived by the defendant.

[3]This assumption allows cost contracts to be possible in that costs are at least partially verifiable, but it puts limitations on what is possible: for example, when adaptation must occur due to changed conditions, the buyer cannot fully determine the exact costs of the modification due to the sellers private information.

[4]Despite the great flexibility provided to the seller by allowing arbitrary choice of (α, β), cost-plus or fixed-price dominate. One possible reason for this regularity is that for any $\beta > 0$, the buyer must expend resources monitoring costs, no matter how small β. The fixed-cost of cost-plus probably introduces a non-convexity that pushes buyers and seller to boundary-point contracts.

adverse weather, technological changes, input price fluctuations, and so on. Assume that there are T possible contingencies for a particular project that can occur, and let π_t be the probability that state $t \in T$ occurs, with the states ordered such that $\pi_t \geq \pi_{t+1}$. The buyer's costs of specifying a contract action for a state is fixed at $k > 0$. The seller's actual costs for adapting the project to meet a contingency is a random variable m distributed according to the commonly known pdf f and cdf F on the support $[0, v - k]$.

Because m does not depend on the particular contingency, the buyer will enumerate states in order, from highest to lowest probability, $1, ...S$, with S being the total number of states enumerated. The buyer's problem is to choose S and (α, β). For simplicity, choosing the optimal S can be interpreted as choosing an optimal probability of not encountering an unenumerated contingency $\tau \in [0, 1]$, and then choosing the smallest S such that $\sum_{t=1}^{S} \pi_t \geq \tau$ as in Equation 1.

$$d(\tau, T) = \min_{S \in \{1,...,T\}} Sk \quad \text{subject to} \quad \sum_{t=1}^{S} \pi_t \geq \tau \tag{1}$$

Renegotiating the Fixed-Price Contract

With probability $1 - \tau$, some state arises that the buyer and seller did not write into the contract and therefore must renegotiate. If the buyer has the bargaining power, then they buyer makes an offer w that maximizes

$$F(w)(v - w) - k$$

where $F(w)$ is the probability that $w > m$, which is the probability that the seller accepts the buyer's offer. If $w < m$, then the seller will walk away from the project. The first-order condition for the seller's maximization problem is

$$w^* = v - \frac{F(w^*)}{f(w^*)} < v$$

For all log-concave density functions, $F(w)/f(w)$ is monotonically increasing, and hence $w^* < v$, which implies that negotiation will fail with positive probability, namely, when $m > w^*$ i.e. when the seller's actual costs are greater than what the buyer is offering. If the seller has the bargaining power, then the seller will simply request v to perform the change and the buyer will always accept the offer. Assuming that sellers do not always have the bargaining power, then some positive proportion of all fixed-price contracts should lead to negotiation failure i.e., litigation.

Renegotiating the Cost-Plus Contract

Under a cost-plus contract, if the buyer has the bargaining power, then the buyer cannot do better than to offer the seller the option to perform according to the unamended contract, but if the seller has the bargaining power, the optimal offer is to perform the work for $v - \mathbb{E}[m]$. In either case, there is no bargaining friction or possibility of negotiation failure, so long as total costs are measurable, even if the cost of modifications is not separately identifiable. One important difference between the two contractual forms is that in the cost-plus form, there is no right to demand contract changes when an unenumerated change occurs, the contract is clear: the seller performs the work at cost and the buyer must reimburse the seller for the realized costs. Bajari and Tadelis nicely summarize the the key difference between the contracts with respect to bargaining:

> In summary, there is a fundamental difference between having a fixed-price or cost-plus contract governing the relationship. A cost-plus contract is a well-defined compensation scheme for both the initial design and any modifications that are requested, so long as compensation is based on total costs. If a fixed-price contract was initially chosen, the compensation scheme is a specific performance compensation scheme and cannot account for modifications, resulting in ex post inefficient bargaining.

The model presented above is greatly simplified and does not include many of the technical points or motivations for the modeling assumptions. However, the two most salient implications from the model are that (a) fixed-price contracts are more likely to lead to bargaining failure, and (b) more complex projects require a larger S to achieve the same τ, all things being equal, and hence the greater design costs Sk of a complex project potentially outweigh the higher-powered incentive benefits of the fixed-price contract. These two testable implications are the focus of the rest of the paper.

3 Data

To create a suitable sample of contracts exhibiting both the fixed-price and the cost-plus structures, I selected contracts issued by the Department of Defense (DoD) during fiscal years 1993 —1996. The dependent variable measuring bargaining failure is an indicator for whether or not a particular contract led to the contract being docketed with the Armed Services Board of Contract Appeals (ASBCA). Using government procurement contracts raises questions about external validity because the government is a unique principal, with motivations, legal constraints, and legal prerogatives that

are different from those of an individual or firm. However, the essential principal-agent dynamic of misaligned interests and differing risk attitudes is maintained in government contracting situations. Once the contract is awarded, the government is concerned with cost and quality and faces the same difficulties in observing effort as a private sector principal.

3.1 Contracts

The Federal Procurement Data System (FPDS)[5] is an on-line database of all contracts issued by the federal government. The FPDS data contain information about the agency issuing the contract, the good or service being procured, the bidding process, the vendor performing the work, and a host of other variables, not all of which are available for every contract. Each observation in the database is a contract "action," such as the initial awarding of the contract, as well as subsequent modifications. Unfortunately, this "action" structure creates one of the greatest challenges in working with FPDS data because it makes it difficult to identify contracting relationships. A single contract might appear numerous times in the database because multiple contract actions might be made under the same contract. Another problem is that even though each contract is assigned a unique contract number, work is often broken up into several "contracts" even if there is only one vendor. Though there are technically multiple so-called contracts in these situations, there is only one contract both legally and in the economic sense of the term. Because the data set is too large for each contract to be investigated individually, some rules were used to infer contract relationships from contract actions, the details of which are in Appendix A and in the commented computer code generating the data set (available from the author's website) from the raw download from the FPDS website.

The population of contracts for this analysis is fully-competed[6] DoD contracts with awards or subsequent modifications greater than $250,000 (in 1996 USD) issued between September 1, 1992, and August 31, 1996, corresponding to fiscal years 1993 through 1996. Though contract data are available up to the present, a more recent time-frame cannot be used because of the slow rate at which disputes arise and are adjudicated. Contracting offices, of which there are several hundred across the Department of Defense, are service-specific (Army, Navy, Air Force, and so on), geographically based offices that actually prepare the contracts, solicit bids, and make awards. As will be discussed the later, the identification strategy relies upon having multiple contracts for the same good or service and multiple procurements across contracting offices. For this reason, the population consists only of contracts for which there were 200 or more procurement contracts for that type of good or service.

[5] https://www.fpds.gov/

[6] These are contracts awarded under an open-bidding process.

This 200 contract cut-off is somewhat arbitrary, so various levels are used in the section covering robustness checks. The results are insensitive to various levels (see Appendix B).

3.2 Disputes

To find which contracts from the population led to disputes, I used a comprehensive listing of the ASBCA's docket from January 1993 to September 2007 obtained through a Freedom of Information Act request. This listing contains all appeals by firms that were filed with the board, even if they were settled between the time of filing and the board hearing.[7] Contracts leading to a dispute have a dependent variable of $disp = 1$ while those that do not have $disp = 0$. Though I use only *disp* as the dependent variable in this analysis, it is possible to determine which of the disputed cases actually led to a board hearing. Contracts leading to a trial have $trial = 1$ while those that did not lead to a trial have $trial = 0$.[8]

Though I do not make use of the *trial* outcomes in this analysis, I do include them in the data set because they allow future researchers to explore the "selection hypothesis," i.e., cases leading to trial are different in important ways from those that settle. This analysis clearly shows that disputed contracts are not simply random samples from the population of all contracts, but the sub-sample of disputed contracts that actually go on to board hearings seems to be random. Figure 1 depicts the changes in means for various binary variables conditional upon whether the contract is from the population, the disputed sub-sample, or the trial sub-sub-sample. The data suggest that selection occurs in the first stage, when contracts lead to disputes, but not in the second, when the parties to disputed contracts decide whether to settle or litigate. Figure 1 also foreshadows the main empirical results: note that fixed-price contracts make up 73% of all contracts in the sample but 93% of disputed contracts.

3.3 Data Description

The most important feature of the contract data set is the type of contract chosen. While there are a wide variety of potential contractual forms available to contracting officers, there are two main types: fixed-price and cost-plus. The Federal Acquisition Regulations (FAR) does not mandate the

[7] One interesting feature of federal government contracting is that each contract includes a "finality clause" that reserves for the government the right to decide on all matters of fact. For this reason, in every case before the ASBCA, the appellant is a private firm (i.e., the seller), and the respondent is the government. For a detailed discussion of the ASBCA, as well as of the history and implications of the finality clause, see Shedd (1964).

[8] To determine which cases actually led to a hearing (which I refer to as a "trial" for simplicity), I used the text of the actual board decisions available from Lexis-Nexis covering the period of January 1, 1993, to the present. Unlike those of other courts, every ASBCA decision leads to a written decision, which allows the creation of a comprehensive listing of all heard cases from Lexis-Nexis data.

use of particular contract forms, but it does contain a discussion of the costs and benefits of different forms. This discussion does not directly present a transaction cost / hold-up argument, but the basic conventional wisdom about the incentivizing effects of fixed-price versus the potentially inefficiency when risks are great is made explicit (see FAR 16.104 - Selecting Contract Types).

For each contract, there is a product or service category code (PSC)[9] that is a fine-grained description of what the contract actually procured. The PSC is used to create an a comprehensive set of indexed dummy variables, *good*. The agency (i.e., Army, Navy, Air Force, and so on) procuring the contract is also represented by a set of indexed dummy variables, *agency*. Each contract issued by the DoD is classified as a definitive contract, a purchase order or a delivery order.[10] This categorical variable is represented by the indexed dummy variables, *type*. The log of the total recorded contract amount is $lamount$.

The good or service, agency, and type variables, along with the continuous contract amount variable, are the independent variables used in the estimating equation. The contract data set contains other variables that might impact ex post bargaining, such as the firm size, level of competition, and so on, but these variables are affected by the choice of $fixed$ and thus are not appropriate controls (nor are they necessary controls). However, some of these variables provide insight into the data and might have future value to other researchers, so I will briefly describe how they were created and what they measure.

I created variables that proxy for the level competition, the nature and cost of the work specified by the contract and various characteristics of the firm. Competitiveness is measured $mbids$, an indicator for whether the contract received two or more bids, and fbo, an indicator for whether the contract was listed as a "Federal Business Opportunity," a government effort to consolidate and publicize federal work on-line and through trade publications. Firm size is measured with an indicator variable *small*, which tells whether the vendor was classified as a small business according to the definition of the Federal Acquisition Regulations. The *special* variable indicates whether the firm was classified as being owned by a woman, veteran, or underrepresented minority. The variables and their definitions are summarized in Table 2. An auxilliary analysis of the effect of $fixed$ on these regressors is included in Appendix C.

[9] The PSC is hierarchical and appears to have a level of specificity similar to that of the NAIC.

[10] A definitive contract is written agreement specifying contingencies and what someone normally thinks of as a contract. A delivery order is a request for an unspecified amount of some good or service (though usually subject to some maximum or minimum amount), while a purchase order is an exact amount procured for an exact price.

3.4 Contract Amount

Figure 3 shows how various regressors change at different contract amount levels. Panel A shows that the proportion of disputed contracts rises with the contract amount, which is unsurprising given the greater stakes and greater potential for disagreement that come with larger contracts. In Panel B, the level of competition, as measured by multiple bids, is falling with larger contracts. This might be due to the small number of firms capable of bidding on the very largest contracts (which also might explain the Panel C results for small businesses), though the trend exists even for smaller contracts.

3.5 Characteristics by Type of Good or Service

Table 3 provides means for various binary regressors crossed with different groupings of good or service. Note that the good-and-service categories are not the exact controls used to create *good* — the categories displayed are aggregated categories, not the lowest level of the hierarchical PSC. One noteworthy feature of the data is that for some categories, all or almost all goods or services are procured with a fixed-price contract. An implication of this result is that one must be careful when inferring the causal effect of switching a contact from cost-plus to fixed-price for goods that, in the sample, were never procured with a cost-plus contract since the estimator only estimates a local average treatment effect.

While I wish to avoid an ad hoc, "just so" explanation from the pattern of results, Table 3 is intriguing: note that "R&D", "Guided Missiles" and "Professional, Administrative and Management Support Services" have the lowest rates of fixed-price contracts and comparatively low dispute rates. The kinds of goods having cost-plus contracts seems consistent with intuitions about the difficulty of writing contracts for certain kinds of tasks. By comparison, consider "subsistence" which is almost entirely procured using fixed-price contracts. The commodity nature subsistence, with its well-defined standards, non-specificity amd low risk make it an ideal cateteory of goods for fixed-price contracts. Perhaps unsurprisingly, no subsistence contracts were disputed in the sample.

4 Empirical Strategy

The empirical goal of this paper is to test the Bajari and Tadelis model by (a) identifying the causal effect of a fixed-price contract structure on the probability of bargaining failure, and (b) comparing that estimate to the naive, endogenous estimate of the effect of the fixed-price structure

on bargaining failure. The empirical strategy is to exploit idiosyncratic variation in how similar goods are procured in various contracting offices. This variation in procurement practices serves as the random assignment that provides the identification of the causal effect. The instrument for the endogenous $fixed$ is the proportion of all goods and services procured using fixed-price contracts by that office, *not including* that particular contract for which the instrument is calculated. This instrument is similar to the one used by Levitt and Snyder in their investigation of how federal spending in a district affects a member of congress's probability of reelection, where they used federal spending in nearby congressional districts as an instrument for spending within a district (Levitt and Snyder Jr, 1997).

4.1 The Reduced-Form Model

Equation (2) is a a reduced-form linear probability model (LPM) for the probability of dispute as a linear function of the characteristics of the contracting relationship, X_i.

$$P(disp_i = 1|X_i) = \beta_0 + \beta_1 lamount_i + \beta_2 fixed_i + \sum_l type_l + \sum_j good_j + \sum_k agency_k + \epsilon_i \qquad (2)$$

Given that *disp* is binary, using a linear probability model (LPM) might seem unattractive, because it can yield probability predictions outside of $[0, 1]$ and is inherently heteroscedastic.[11] However, with endogenous binary regressors, using the LPM is a relatively simple strategy with directly interpretable coefficients, and in practice the model seems to generate marginal effects estimates quite similar to those produced with more complex procedures (Angrist, 2001). Furthermore, because the main goal of this analysis is to identify marginal effects and not actual predicted probabilities, the possibility that the LPM will predict probabilities outside $[0, 1]$ is relatively unimportant.

The most important argument in favor the LPM in this particular application is the presence of a large number of group-indicating dummy variables. In a general linear model (GLM), the coefficients on dummy variables for groups that are homogeneous in their outcomes (i.e., all disputed or none disputed) are not identified (Caudill, 1987, 1988). Estimating GLMs like probit or logit with maximum likelihood methods will require an infinitely positive coefficient on "always" groups and an infinitely negative coefficient on "never" groups. Lastly, for rare-event data such as contract litigation, it is not clear that either probit or logit is a particularly good choice (King and Zeng, 2001).

[11] This difficulty can be easily corrected with the use of robust standard errors.

4.2 Identification Assumptions

Since $fixed$ is endogenous, β in Equation 2 is unidentified. I will present an instrumental variables approach that will allow identification, but before doing this, it is necessary to discuss the general assumptions that will provide identification of β in Equation 2:

1. The government's decision to procure the good or service and the specific agency procuring the good or service is exogenously determined. The basic contractual form (but not the incentive structure), such as whether the government procures the good with a delivery order, purchase order or a definitive contract, is also exogenous.

2. The contract amount, conditional upon the other regressors, is exogenous.

3. The contracting office has an impact on various contract characteristics but the contracting office has no direct impact on the probability of dispute or litigation, though I do allow and control for agency fixed-effects. This assumption is essentially the exclusion restriction for an instrumental variables estimator.

4. Unobserved contract characteristics are uncorrelated across contracts within a contracting office, conditional upon the good or service being procured and the approximate contract amount. This assumption must be strengthened to independence if $P(disp_i = 1|X_i)$ is modeled as some form of GLM, though this is not necessary in the LPM case.

4.3 Justification of the Assumptions

Exogeneity of Government Need and Contract Amount

I assume that the government requirement that generates the contracting relationship is exogenous and that any regressors that are determined by that need are as exogenous as well. This assumption is quite plausible because government needs are not determined or generated by contracting offices — they merely implement decisions made by other government bureaucracies. The assumption that recorded contract cost is an unbiased estimate of the true economic cost of a proposed project is controversial because the recorded contract amount is determined in part by the bidding firm. However, by design, every contract in the sample was awarded through a competitive process. Though the government might be expected to get a better or worse deal on any particular contract, it seems unlikely that the contract amount, conditional upon the other regressors, is chosen by anyone or is correlated with factors beyond simply what is being procured.

The Effect of the Contracting Office

Every government contract is issued under the authority of a single contracting officer, who is a warranted agent authorized to commit the government to a contract. Given the high potential for misconduct, contracting officers have fairly limited discretion and their decisions are subject to scrutiny by third parties. Despite these measures, contracting offices do exhibit idiosyncratic preferences and customs. The existence of these office effects seems to contradict the identification assumption that contracting offices have no direct impact on litigation except through their effect on contract structure. However, two stylized facts about federal procurement resolve this apparent violation of the exclusion restriction.

First, even though all contracts must be issued by a contracting office, the day-to-day management rarely comes from the contracting office; rather it comes from the agency that a contracting office serves. For example, a contracting office might prepare the written contract for the construction of new barracks at an Army base, but the day-to-day principal-agent interaction would come from the government's project manager, not from the contracting officer that drafted the contract. Because the dispute is likely to arise externally to the contracting office, the exclusion restriction assumption seems more plausible.

Second, the government's decisions regarding a contract reaching the ASBCA will almost certainly be made by government lawyers or higher-level executives and not by the contracting officer (and recall that the reduced-form equation includes controls for the agency providing the lawyers or making settlement decisions). Therefore, the ultimate decision on whether to accept or reject an offer is not made by the contracting office or the contracting officer responsible for the incentive structure of the contract: ex post negotiation and contract-drafting are compartmentalized in such a way that the contract writers have no direct effect on the probability of litigation once the project is underway.

4.4 Creating an Instrument from Idiosyncratic Contract Office Effects

If there are contract office customs that influence the decision to prefer one contract form over another, then a natural instrument for $fixed$ would be some measure of how that office procured similar goods or services in other contracting situations. If it is assumed that unobserved contract characteristics are independent within an office, then any instrument generated from other contracts procured by that office should also be uncorrelated with the error term for any particular contracting situation.

To demonstrate more precisely the problematic endogeneity of $fixed$, consider some contract i of good-and-service category k (such that $good_k = 1$), with a value within some specified range, issued by some particular contracting office. For argument's sake, assume that the contracting officer's decision about which structure to choose for each contract can be modeled with a simple LPM such as Equation 3, where α_g captures the office and good-and-service fixed effects and ν_i is a composite error term capturing all unobserved effects on the choice for $fixed_i$.

$$P(fixed_i = 1) = \alpha_g + \nu_i \tag{3}$$

By construction, $\mathbb{E}[\nu_i] = 0$ and $\mathbb{E}[\alpha_g \nu_i] = 0$. The unobserved component ν_i might include the unobserved task complexity, and presumably this task complexity also can affect $P(disp_i = 1|X_i)$ directly. Hence, $\mathbb{E}[\epsilon_i' \nu_i] \neq 0$ and therefore $\mathbb{E}[fixed_i' \epsilon_i] \neq 0$. The exogeneity assumption for OLS is thus violated and therefore the effect of $fixed$ on $P(disp_i = 1|X_i)$ is unidentified. The problem of endogeneity is a direct implication of the Bajari and Tadelis model: the complexity that reduces the the contracting officer's incentive to use a fixed-price contract *comes* from the fact that this complexity will increase transaction costs because unanticipated ex post changes will be more likely. The contracting officer can partially address the ex post bargaining problem by writing a more complete contract, but ceteris paribus, the fixed-price contract is relatively less desirable. One method of solving this problem is to find an instrumental variable that is correlated with $fixed$ but uncorrelated with ν and uncorrelated with ϵ.

Returning to the hypothetical grouping of contracts that generated the example of endogeneity, consider Equation 4, which is the proportion of fixed-price contracts in some group (with all contracts having the same *good* and $office$ values), not including the ith contract itself. This z_i will serve as the instrument for the ith contract.

$$z_i = \frac{1}{m-1} \sum_{j=1|j\neq i}^{m} fixed_i \tag{4}$$

Using the Equation 3, z_i can be represented as:

$$z_i = \alpha_g + \frac{1}{m-1} \sum_{j=1|j\neq i}^{m} \nu_j$$

The first requirement of the instrument is that $\text{corr}[fixed_i' z_i] \neq 0$, which can be verified with the first-stage of our two-stage least squares. The second requirement is that the instrument is uncorrelated

with our two unobserved error components ν and ϵ. For ν, note that:

$$\mathbb{E}[z_i \nu_i] = \mathbb{E}\left[\sum_{j=1|j\neq i}^{m} (\alpha_g + \nu_j)\nu_i\right] = \sum_{j=1|j\neq i}^{m} \mathbb{E}[\nu_j \nu_i]$$

By our assumption that that unobserved characteristics are not correlated across contracts within a specified grouping, $\mathbb{E}[\nu_j \nu_i] = 0$ and hence $\mathbb{E}[z_i \nu_i] = 0$.

For ϵ, by our assumption that there are no contracting office effects and the inclusion of goods-and-services controls in LPM for $P(disp_i = 1|X_i)$, $\mathbb{E}[\alpha_g' \epsilon] = 0$ for all g since α_g is not correlated with any unobserved effects. Our assumption that unobserved contract characteristics are not correlated even within a grouping g also implies that ϵ_i is not correlated with ν_j, where ν_j is the unobserved characteristics in some other random contract from the grouping, or:

$$\mathbb{E}[z_i \epsilon_i] = \mathbb{E}\left[\sum_{j=1|j\neq i}^{m} (\alpha_0 + \nu_j)\epsilon_i\right] = \sum_{j=1|j\neq i}^{m} \mathbb{E}[\nu_j \epsilon_i] = 0$$

To put things more succinctly, by the exclusion restriction assumption, we have $\mathbb{E}[z_i \epsilon_i] = 0$.

To provide some intuition for the instrument, consider that z_i is the empirically determined probability that a given contract in the specified grouping will be selected for $fixed = 1$, but this estimate is not based on unobserved characteristics that might affect the decision in the ith particular contracting situation. The peculiarities of the contracting office, as measured by this instrument, are serving to randomly assign some contracts fixed-price structures, while other offices are assigning comparable contracts cost-plus contracts, but without that office having an independent impact on the probability of a dispute, thus allowing the causal effect of that structure to be identified.

One important point is that this grouping is not necessarily *all* contracts issued by the office: we can further restrict this grouping based on other factors. In all the regressions, I sub-divide contracts based on the median amount for all contracts in the sample. The purpose of this further refinement is that even with controls for the contract amount in the first-stage, there might be fundamental differences between the largest and smallest contracts issued by an office not captured by *lamount*. Obviously, there is a trade-off between making ever smaller groups (and thus hopefully reducing bias or the possibility of mis-specificaton) and eventually creating too many useless singleton groups.

Causal IV Assumptions

A key contention of this paper is that the choice of contract structure has a causal effect on the probability of a dispute arising. For a causal interpretation to be plausible in an IV context, there are several assumptions that must be met, most of which map directly from our standard identification assumptions. Angrist et al. (1996) lay out the necessary assumptions using the potential outcomes framework for causal IV interpretations:

1. *Stable Unit Treatment Value Assumption (SUTVA):* Potential outcomes for each contract i are unrelated to how other contracts are procured. This assumption is almost certainly satisfied, because it seems very unlikely that a firm's decision on whether or not to appeal a contracting officer's decision (and hence set $disp = 1$) depends at all on how any other contracts are procured or on their litigation outcomes (especially considering the often decades-long lag between disputes and board hearings).

2. *Ignorable Treatment Assignment:* The ignorable treatment assumption is clearly not met with a simple OLS approach: presumably those contracts that are selected for a fixed-price contract are selected on the basis of other unobserved (from our perspective) characteristics as well as unobserved characteristics. The purpose of the instrument and the TSLS estimation is to provide the exogenous variation in $fixed$ that replicates the logic of a true experiment when treatment assignment is, by design, random or at least ignorable.

3. *Exclusion Restriction:* The effect of z_i on $disp$ occurs only by affecting $fixed$. This assumption is equivalent to the assumption made here that contracting offices have no direct effect on the probability of litigation, *except* through their effect on $fixed$ (and any effect that change in $fixed$ might have on other covariates). Unfortunately, this is not a testable assumption, but for the peculiar structural reasons discussed earlier (and the use of agency dummies), it is likely met.

4. *Non-Zero Average Causal Effect* This is simply the requirement that the instrument is sufficiently correlated with the endogenous regressor, which is shown convincingly with the first stage of the TSLS regression.

5. *Monotonicity:* This assumption requires only that z_i have a monotonic effect on $fixed$. In other words, having a large number of "neighboring" contracts procured in a similar manner does not make it less likely that the contract in question will also be procured with this method,

which, given the context, seems fairly innocuous. From a bureaucratic perspective, it is almost certainly easier to procure all items with an identical contract because re-using contracts is far cheaper than creating new contracts "from scratch" which makes the monotonicity assumption even more plausible.

5 Results

Table 4 shows the OLS and TSLS, including the first stage. Unsurprisingly, the contract amount has a strong, positive affect on the probability of dispute. Larger contracts simply have greater stakes for the parties and usually are more complex and hence provide more opportunities for disputes. Both the OLS and TSLS estimates have very low R^2 values, which is expected given the fairly small number of disputed contracts and the general nature of the regressors. The first stage of the TSLS shows that our instrument is highly correlated with $fixed$ and that the F-statistic for this first stage is very high.

In the TSLS estimate, using a fixed-price contract causes an increase in the probability of a dispute. The TSLS estimate of the coefficient of $fixed$ is almost twice as large as the OLS estimate. If the Bajari and Tadelis model is correct, the unobserved characteristics that are biasing downward the OLS results are unobserved characteristics such as the cost of writing the contract, the completeness of the design and probability of important changes. While the results do show that a forward-looking appraisal of likely transaction costs do influence the choice of structure, the results do not conclusively show that asymmetry about adaptation costs are the prime reason for ex post bargaining friction.

5.1 Alternative TCE Mechanisms

The Bajari and Tadelis model explicitly argues that it is ex post asymmetry of information about adaptation costs are the cause of bargaining friction. It is possible that the fixed-price structure causes more bargaining friction for non-adaption costs reasons and hence other theories could explain the empirical results without the Bajari and Tadelis model being valid. I have identified four possible ways in which the incentive structure could increase ex post transaction costs unique from the Bajari and Tadelis model:

1. The incentive structure can encourage parties to take unilaterally self-benefiting actions detrimental to the other party. For example, a fixed-price contract creates an incentive for the

seller to reduce costs by shading on quality, while a cost-plus contract creates an incentive to inflate costs. Under this scenario, the fixed-price contract simply creates more violations of the contract terms and hence more litigation results.

2. If the contract allocates greater financial risk to one party, then that party will have an incentive to seek a legal pretext for nullifying the contract when some adverse state of the world is revealed. For example, insurers have an incentive to find policy-voiding errors when the policy holder seeks a payout, but not before the loss-causing contingency occurs. Under this scenario, one or both parties would have to find contract nullification more attractive when the relationship is governed by a fixed-price contract.

3. If a certain incentive structure increases the variance in one party's estimate of the other party's effort, cost reduction or compliance with contract terms or any other salient feature that might be relevant to the court, this greater uncertainty might manifest itself in higher dispute rates. For example, a long, confusing contract with multiple terms is likely to be misinterpreted and hence likely to lead to more error in each parties subjective perception of the other sides compliance with the contract. Under this scenario, the fixed-price contract would have to generate more confusion among the parties about each sides' compliance with the contract terms.

5.2 Evaluation

The first mechanism — more cheating or self-dealing with fixed-price contracts— seems unlikely to be the cause of greater fixed-price litigation because the cost-plus contract is probably more conducive to cheating. First, profit increases and cost padding (i.e. over-reporting costs) are not just correlated in the cost-plus case: they are co-linear. Every dollar padded is a dollar earned and further, such padding poses little risk because the worst-case scenario (absent punitive damages or some kind of fraud liability) is that the seller must give the buyer a rebate. In contrast, shading on effort in a fixed-price contract is only correlated with profit, can lead to long-term, lingering liabilities if quality defects are detected later and can have severe reputation repercussions. Of course, cost-padding also hurts a firms reputation, but most firms firms would rather be known as "good but expensive" than "shoddy but cheap" because presumably a reputation for bad quality is far stickier than an easily-alterable reputation for high pricing.

The second mechanism — more risk in a fixed-price contract — is an unsatisfying explanation because all projects have risk: the contractual form just specifies which party is bearing the prepon-

derance of that risk. No matter which state of the world is revealed, one party or the other has an incentive to obtain better terms. If the project turns out to be difficult and costly, the buyer wishes there had been a fixed-price contract and the seller wishes there had been a cost-plus contract; if the project is easy and not costly, then each side wishes the opposite. This risk-symmetry suggests that greater litigation in the fixed-price case for risk-sharing reasons is not very plausible.

The third mechanism — greater uncertainty regarding performance in a fixed-price contract — seems plausible because detecting fixed-price quality shading is probably harder than detecting cost-plus padding. It may take years for quality problems with some goods to become fully apparent. Even if it is assumed that courts will punish sellers for quality problems (which may have a large stochastic component), this task is daunting in the fixed-price scenario because it requires assessment of the likelihood of both what occurred and what did not occur. By contrast, in a cost-plus situation the problem is comparatively simple: everyone knows what work was done and the reported cost of that work (and unlike in the fixed-price scenario, there is no presumption that the seller might have shaded on quality, because the seller has no incentive to do so). It might be the case that there is simply a greater perceived uncertainty in fixed-price contracts and hence greater divergence in each side's perceived expectation, which in turn decreases the probability of successful bargaining. This question seems interesting and cannot be answered without further study.

5.3 Content Analysis of Legal Decisions

The preceding argument is rather unsatisfying as it appeals to (informal) theory to answer questions that are, at least in principle, empirically answerable. A better approach would be a careful case study of a large number of disputes to see whether bargaining over adaptation costs was the dominant source of friction. Although this may be a fruitful topic for future research, obtaining a sufficient sample size would require a serious effort and is beyond the scope of this paper.

A partial substitute for a case study is a content analysis of the ASBCA Judge's written decisions. If we continually find references to certain keywords indicating haggling over adaptation costs in the text of the decision, it might indicate the flavor of dispute between the government and the firm. Of course, the frequency with which a "bargaining word" appears in the corpus of decisions must be compared to some other body of similar text to determine whether or not a particular word is appearing more or less frequently that we would expect. This method of course does not definitively answer anything and is merely suggestive, it does potentially provide some insight in into the nature of the disputes that might otherwise be completely obscure.

For this analysis, I used all the written decisions of the ASBCA from 1993 to the present as my corpus. For a "control" corpus, I use the text of decisions by the Occupational Safety & Health Review Commission Administrative Law Judge from January 1993 to February 1993.[12] I chose this corpus because it has some similarities to the ASBCA text. Like the ASBCA decisions, the written decisions from the OSHR tend be fact-oriented, with a focus on whether a firm was in violation of some particular OSHA regulation. If we imagine that most contract disputes are not about adaptation costs but rather the firm's compliance with an as-if complete contract, then the two court's tasks, and hence decisions, might be similar.

Figure 5.4 shows a log-log plot of word frequencies in the ASBCA decision corpus (y-axis) and the OSHR decision corpus (x-axis). A selection from the one-hundred most frequent English words were used as well as a small number of "bargaining words" such as "offer" and "reject" and "adaptation words" such as "design", "delays", "changes" etc. Words above the 45 degree line in Figure 5.4 are found more frequently in the ASBCA corpus, while those on the line are roughly equi-probable. Notice that "changes" occurs approximately eight times more frequently in the ASBCA corpus, "modifications" almost twice as often and "design" six times as often. The bargaining words — "reject" and "offer" also appear far more frequently in the ASBCA corpus. While these results are strongly caveated, they are suggestive and a further analysis seems likely to confirm the essentials of the Bajari and Tadelis argument about the nature of the disputes.

5.4 Conclusion

The main empirical finding of this paper is that the use of a fixed-price contract increases, on average, the probability of bargaining friction and subsequent litigation. The paper also suggests that contracting officers are aware of the trade-offs between the incentives for cost reduction with the fixed-price contract and the ex post bargaining friction of such contracts. A content analysis of the appearance of certain words in the text of ASBCA decisions compared to a control corpus suggests that bargaining and adaptation are at the heart of most contract disputes. The empirical results and the content analysis are consistent with the procurement model put forth by Bajari and Tadelis and suggests that ex post renegotiation and transaction costs are a significant feature of the contracting problem.

[12] Even with this much smaller time frame, the OSHR corpus provides over 600,000 words - far more than the ASBCA sample of 200,000 words.

Table 1: Comparing Fixed-Price and Cost-Plus Contracts

Features	Fixed-Price	Cost-Plus
Risk allocation mainly on	Seller	Buyer
Incentives for quality	Less	More
Buyer administration	Less	More
Good to minimize	Costs	Schedule
Documentation efforts	More	Less
Flexibility for change	Less	More
Adversarial Relationship	More	Less

* Reproduced from Bajari and Tadelis (2001)

Table 2: Some Key Contract Variables

Name	Description
$l/amount$	log / contract amount
$good$	controls for good/service being procured
$fixed$	indicator for fixed-price (not cost)
$mbids$	indicator for # bids ≥ 2
$small$	indicator small business
$special$	indicator for woman or minority-owned

Table 3: Contract Characteristics by Type of Good or Service

Good	N	disp	fixedprice	small	special	mbids	fbo	avg. amt
All	40417	0.02	0.73	0.46	0.03	0.3	0.7	4.04
ADP EQUIPMENT, SOFTWARE, SUPPLIES AND SUPPORT EQUIPMENT	1645	0.03	0.941	0.47	0.046	0.332	0.595	1.84
AIRCRAFT AND AIRFRAME STRUCTURAL COMPONENTS	793	0.02	0.878	0.311	0.004	0.564	0.753	30.65
AIRCRAFT COMPONENTS AND ACCESSORIES	337	0.006	0.899	0.291	0.006	0.665	0.792	5.68
AMMUNITION AND EXPLOSIVES	267	0.011	0.82	0.217	0	0.539	0.757	0.69
ARCHITECT AND ENGINEERING SERVICES - CONSTRUCTION	1395	0.009	0.884	0.238	0.011	0.091	0.939	2.56
AUTOMATIC DATA PROCESSING AND TELECOMMUNICATION SERVICES	1368	0.007	0.513	0.45	0.072	0.482	0.614	5.42
CLOTHING, INDIVIDUAL EQUIPMENT, AND INSIGNIA	315	0.006	0.902	0.46	0	0.324	0.508	3.37
COMMUNICATION, DETECTION, AND COHERENT RADIATION EQUIPMENT	978	0.016	0.817	0.232	0.006	0.517	0.71	6.21
CONSTRUCTION OF STRUCTURES AND FACILITIES	4119	0.039	0.976	0.511	0.028	0.092	0.806	3.12
EDUCATION AND TRAINING SERVICES	367	0	0.992	0.008	0	0.826	0.016	0.67
ELECTRICAL AND ELECTRONIC EQUIPMENT COMPONENTS	550	0.007	0.898	0.171	0.007	0.665	0.82	2.3
ENGINES, TURBINES, AND COMPONENTS	910	0.007	0.945	0.335	0	0.405	0.766	8.18
FUELS, LUBRICANTS, OILS, AND WAXES	1456	0.004	0.994	0.527	0.012	0.152	0.944	11.07
FURNITURE	370	0	0.786	0.143	0.003	0.032	0.408	0.45
GUIDED MISSILES	198	0	0.495	0.045	0	0.611	0.672	14.61
INSTRUMENTS AND LABORATORY EQUIPMENT	399	0.003	0.935	0.316	0.005	0.424	0.82	1.48
MAINTENANCE AND REPAIR SHOP EQUIPMENT	193	0.016	0.839	0.415	0	0.575	0.751	2.31
MAINTENANCE, REPAIR, AND REBUILDING OF EQUIPMENT	1461	0.023	0.772	0.346	0.027	0.344	0.737	5.57
MAINTENANCE, REPAIR OR ALTERATION OF REAL PROPERTY	8331	0.031	0.985	0.706	0.042	0.131	0.761	1.41
MISCELLANEOUS	149	0.007	0.537	0.463	0.027	0.57	0.356	5.46
PROFESSIONAL, ADMINISTRATIVE AND MANAGEMENT SUPPORT SERVICES	3348	0.008	0.24	0.352	0.036	0.465	0.728	5.61
R&D	6436	0.003	0.089	0.41	0.01	0.3	0.813	3.43
SUBSISTENCE	2222	0	0.999	0.404	0.014	0.588	0.178	1.42
TRANSPORTATION, TRAVEL AND RELOCATION SERVICES	338	0.009	1	0.216	0.003	0.071	0.325	9.95
UTILITIES AND HOUSEKEEPING SERVICES	2472	0.028	0.958	0.523	0.089	0.343	0.409	2.28

Notes: ADP: Automatic Data Processing disp: fraction of contracts disputed
fixedprice: fraction that are fixed-price
fbo: fraction awarded through the Federal Business Opportunities program
small: fraction that are small businesses
special: fraction of contracts awarded to woman, minority or veteran-owned businesses
avg. amount: average contract amount, in millions of 1996 USD.

Table 4: Effects of Contract Characteristics on Probability of Dispute

Variables	OLS	First Stage	TSLS
Log Amount	0.00509	-0.00937	0.00528
	(0.00064)	(0.00114)	(0.00065)
Fixed-Price	0.0129		0.02188
	(0.00197)		(0.00458)
Fixed-Price Instrument		0.69445	
		(0.01099)	
N	40417	40417	40417
R-Squared	0.0255	0.7339	0.0253
F-statistic	9.4215	992.4058	9.34

Notes — The dependent variable is $P(disp_i = 1|X_i)$. The standard errors are robust. Both regressions include fixed effects for the agency, good or service and sub-contract plan. The two-stage least-squares (TSLS) estimate includes a single instrument, z_i, for the endogenous binary regressor $fixed$.

Figure 1: Comparison of Population, Disputed and Litigated Contracts

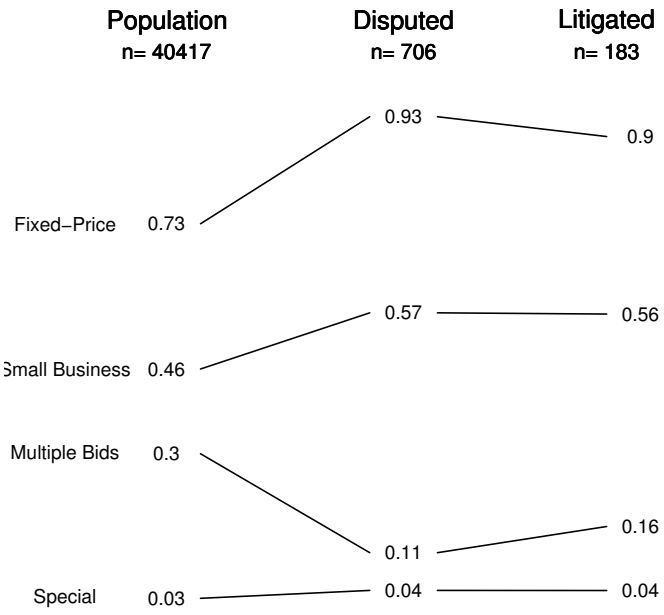


Figure 2: Trends in various Regressors and Outcomes vs. Contract Amount

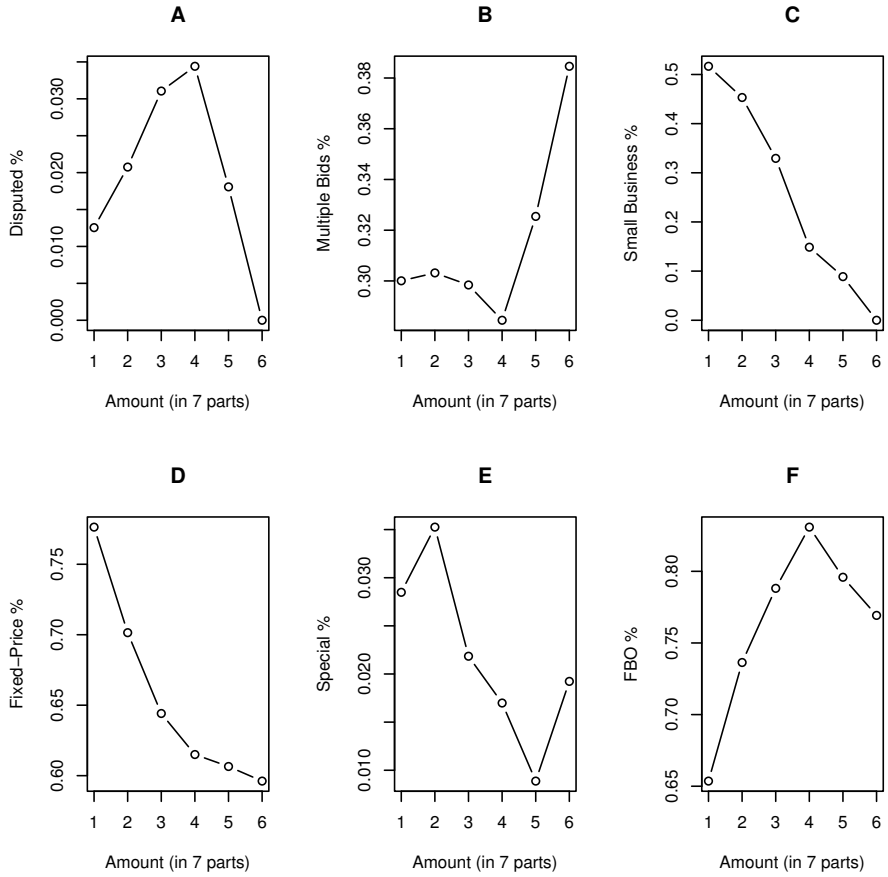
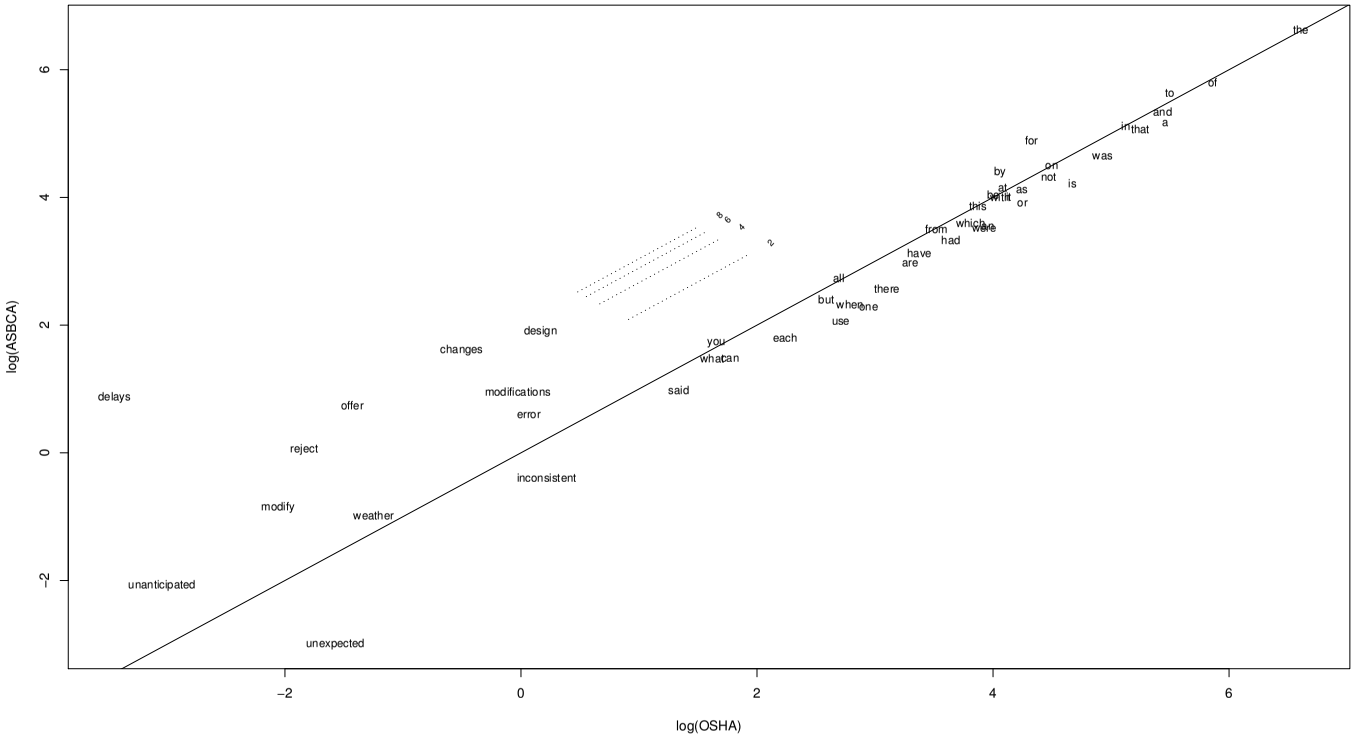


Figure 3: Comparison of Word Frequencies in ASBCA Decisions versus OSHA Decisions



This plot compares the log of word frequencies (appearances / 10,000 words) for the OSHA corpus on the x-axis and for the ASBCA corpus on the y-axis. The exact point on the plot is centered in each word, both in word length and font height. The dashed lines indicate iso-multiple lines: for example, words on the dashed line labeled 2 (above the 45 degree solid line) appear twice as frequently in the ASBCA corpus as in the OSHA corpus. Note the much higher frequency of "bargaining words" or "adaptation words", which is evidence in favor the Bajari and Tadelis mechanism.

References

Angrist, J., G. Imbens, and D. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association 91*(434).

Angrist, J. D. (2001, jan). Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business and Economic Statistics 19*(1), 2–16.

Bajari, P. and S. Tadelis (2001). Incentives versus transaction costs: A theory of procurement contracts. *The RAND Journal of Economics 32*(3), 387–407.

Banerjee, A. V. and E. Duflo (2000, aug). Reputation effects and the limits of contracting: A study of the Indian software industry. *The Quarterly Journal of Economics 115*(3), 989–1017.

Bebchuk, L. A. (1984). Litigation and settlement under imperfect information. *The RAND Journal of Economics 15*(3), 404–415.

Caudill, S. B. (1987). Dichotomous choice models and dummy variables. *The Statistician 36*(4), 381–383.

Caudill, S. B. (1988). An advantage of the linear probability model over probit or logit. *Oxford Bulletin of Economics and Statistics 50*(4), p425 – 427.

Eggleston, K., E. Posner, and R. Zeckhauser (2000). The design and interpretation of contracts: Why complexity matters. *Northwestern University Law Review 95*(1), 91–132.

Eisenberg, T. (1990, jun). Testing the selection effect: A new theoretical framework with empirical tests. *The Journal of Legal Studies 19*(2), 337–358.

Hylton, K. N. (1993, jan). Asymmetric information and the selection of disputes for litigation. *The Journal of Legal Studies 22*(1), 187–210.

Kessler, D., T. Meites, and G. Miller (1996, jan). Explaining deviations from the fifty-percent rule: A multimodal approach to the selection of cases for litigation. *The Journal of Legal Studies 25*(1), 233–259.

King, G. and L. Zeng (2001). Logistic Regression in Rare Events Data. *Political Analysis 9*(2), 137.

Laffont, J. and J. Tirole (1993). *A Theory of Incentives in Procurement and Regulation*. MIT Press.

Levitt, S. and J. Snyder Jr (1997). The Impact of Federal Spending on House Election Outcomes. *The Journal of Political Economy 105*(1), 30–53.

Priest, G. and B. Klein (1984, jan). The selection of disputes for litigation. *The Journal of Legal Studies 13*(1), 1–55.

Shavell, S. (1996, jun). Any frequency of plaintiff victory at trial is possible. *The Journal of Legal Studies 25*(2), 493–501.

Shedd, J. P. J. (1964). Disputes and appeals: The armed services board of contract appeals. *Law and Contemporary Problems 29*(1), 39–86.

Spier, K. E. (2008). Economics of litigation. In S. Durlauf and L. Blume (Eds.), *The New Palgrave Dictionary of Economics (forthcoming)* (2 ed.). Palgrave Macmillian.

Whinston, M. (2001). Assessing the Property Rights and Transaction-Cost Theories of Firm Scope. *The American Economic Review 91*(2), 184–188.

A Data Generation

As noted in Section 3, one of the difficulties of working with FPDS data is that contracts are stored as "actions," where an action might include both the original awarding of the contract as well as subsequent modifications. And additional complication is that some very large projects are broken into several contracts, though for the purposes of this paper, all these sub-contracts are really part of one contracting relationship. Identifying these relationships is a challenge and requires some inferences and imputations. The actual raw data and the code used to generate the data set are available on the author's website, but for clarity, the process by which the data was constructed and shaped is outlined here. Starting with the raw set of contracts:

1. Contracts with identical contract numbers issued by the same contract office are treated as one contract.

2. If there is any disagreement among the categorical covariates (i.e. one action records a fixed-price contract while another records a cost-contract), options are weighted by the contract amount and the option with the largest share of the total contract amount is recorded as the actual covariate for the amalgamated contract.

3. Contracts with missing contract numbers, contract offices and goods and service categories are dropped from the sample.

B Robustness

Given that the data set is somewhat unusual, with a large sample size but exceedingly small response rate, there might be concerns that the estimation strategy is incapable of recovering the LATE or that standard error approximations are incorrect. Further, as with almost all instrumental variables estimation strategies, there might be concern about the strength of the instrument and whether the exclusion restriction truly holds. Lastly, there might be concern that the pattern of results — never-mind the parameter estimates and standard errors — are sensitive to the functional form or the necessary but somewhat arbitrary decisions about which observations to include in the sample. Although some of the underlying assumptions are ultimately untestable, others can be made more plausible by subjecting them to appropriate falsification tests.

B.1 Alternate Specifications, Alternate Instrument Formulations

The paucity of non-dummy exogenous variables makes it difficult to develop meaningful alternate specifications: too many dummy variable interaction terms and the covariate space quickly becomes partitioned into spaces of perfect prediction. One fairly simple check is to use levels instead of logs for the contract amount. Table 5 shows the results from Table 4, but with levels instead of logs. Notice that the basic pattern of results does not change, though the magnitudes are decreased somewhat comapred to the log case.

B.2 Pseudo-Instruments & Pseudo-outcomes

The identification strategy depends critically on contract offices exerting an influence on the choice of contract structure, but not directly on the probability of litigation. While we cannot test this assumption directly, we can investigate several counter-factuals that might illuminate whether the office is truly having an effect on $fixed$. One method for investigating this assumption is use a a pseudo-instrument, z_i^p, based on the randomization of contracting office assignment. If there is no persistent contracting office effect on $fixed$, the pseudo-instrument, while still correlated with $fixed$, should have an uninformative, "weak instrument" first stage with a statistically insignificant coefficient on the predicted $fixed$ in the second stage. If we obtain the same results with the pseudo-instrument as with the real-instrument, we have found a serious problem: some kind of

Table 5: Effects of Contract Characteristics on Probability of Dispute - Levels instead of Logs

Variables	OLS	First Stage	TSLS
Amount (10 millions)	0	-0.00144	1e-05
	(1e-04)	(0.00063)	(1e-04)
Fixed-Price	0.01069		0.01653
	(0.00193)		(0.00447)
Fixed-Price Instrument		0.70063	
		(0.01095)	
N	40417	40417	40417
R-Squared	0.0234	0.7334	0.0232
F-statistic	8.6173	989.8286	8.5435

Notes: The dependent variable is $P(disp_i = 1|X_i)$. The standard errors are robust. Both regressions include fixed effects for the agency, good or service type and contract type (i.e., delivery order, puchase order or definitive contract). The two-stage least-squares (TSLS) estimate includes a single instrument, z_i, for the endogenous binary regressor $fixed$. This table shows that the basic pattern of results does not change when the regression is done in contract amount levels intead of in logs.

mechanical correlation or mis-specification is driving the TSLS results rather than the proposed causal mechanism.

Table 6 replicates Table 4 but using a pseudo-instrument generated from randomly assigning contracts within a good or service category to different offices. From the first-stage it is obvious that this instrument has a serious weak instruments problem and the TSLS estimate, as expected, gives non-sensical results.

B.3 Variations on Quasi-Arbitrary Cut-offs

In the results presented in the main body of the paper, only contracts for which there were 200 individual contracts for that item were included in the analysis. Furthermore, only those contracting offices with at least 15 procurements in the sampled period were included in the sample. Table 7 shows TSLS results with various contract number cut-offs but still with 15 required contracts per office, while Table 8 shows results with a minimum of 30 contracts per office. While the estimated LATE is sensitive to the choice of cut-offs, the basic pattern of results is maintained, with a positive sign on $fixed$ and a downward bias in the OLS estimate.

Table 6: Replication of Table 4 with Pseudo-Instrument

Variables	OLS	First Stage	TSLS
Log Amount	0.00509	-0.02104	0.01758
	(0.00064)	(0.00134)	(0.00955)
Fixed-Price	0.0129		0.60621
	(0.00197)		(0.45339)
Fixed-Price Instrument		-0.00363	
		(0.00323)	
N	40417	40417	40417
R-Squared	0.0255	0.6523	0.0249
F-statistic	9.4215	675.112	9.1883

Notes: The dependent variable is $P(disp_i = 1|X_i)$. The standard errors are robust. Both regressions include fixed effects for the agency, good or service and sub-contract plan. The two-stage least-squares (TSLS) uses single instrument for $fixed$, z_i^p, a pseudo-instrument generated by randomizing contract office assignments.

Table 7: Fixed-Price Effects with Variation in Contract Cut-off Amounts: 15 Contracts Per Office Minimum

V1	100	250	300
OLS	0.01396	0.01299	0.01285
	(0.00174)	(0.00194)	(0.00206)
TSLS	0.02229	0.02153	0.02127
	(0.00416)	(0.00461)	(0.00489)

Notes: The dependent variable is $P(disp_i = 1|X_i)$. The standard errors are robust. All regressions include fixed effects for the agency, good or service type and contract type (i.e., delivery order, puchase order or definitive contract). The two-stage least-squares (TSLS) estimate includes a single instrument, z_i, for the endogenous binary regressor $fixed$.

Table 8: Fixed-Price Effects with Variation in Contract Cut-off Amounts: 30 Contracts Per Office Minimum

V1	100	250	300
OLS	0.01481	0.0141	0.01417
	(0.00183)	(0.00209)	(0.00221)
TSLS	0.02635	0.02556	0.02521
	(0.00445)	(0.005)	(0.00523)

Notes: The dependent variable is $P(disp_i = 1|X_i)$. The standard errors are robust. Both regressions include fixed effects for the agency, good or service type and contract type (i.e., delivery order, puchase order or definitive contract). The two-stage least-squares (TSLS) estimate includes a single instrument, z_i, for the endogenous binary regressor $fixed$.

C Other Outcomes

Equation 2 did not include many observed variables available from FPDS because they clearly are not exogenous. For example, the firm size is recorded for each seller, but because small firms might choose whether or not to bid on work based on the incentive structure, a small business indicator is really an outcome variable, not an independent, explanatory variable.

Table 9 shows the average causal effect of $fixed$ on the probability of the contract (a) being awarded to small business (b) receiving multiple bids (c) being awarded to a woman, minority of veteran-owned business and d) being awarded to a firm that had multiple contracts with the government during the sampled period. These results are generally consistent with economic intuition: smaller firms are less likely to bid on the largest projects, which explains the negative sign on $lamount$, and are probably less capable of bearing the risk associated with fixed-price contracts. A similar line of reasoning can explain the pattern of results for the "special", (c) regression. The effect of fixed-price on receiving multiple bids is most likely a result of the government rationally attempting to publicize fixed-price contract opportunities and increase the number of bidders, because having a large bidding pool is likely to reduce costs in a fixed-price setting. That $fixed$ decreases bidding by firms with multiple government contracts might be a residual effect of $fixed$ increasing the number of bids and thus attracting a more heterogeneous group of firms.

Table 9: Effects of Contract Characteristics on Other Outcomes (TSLS)

Variables	Small Business	Multiple Bids	Special
Log Amount	-0.05617	-0.00652	-0.00125
	(0.0018)	(0.00177)	(0.00065)
Fixed-Price	0.06908	-0.03143	-0.01723
	(0.0177)	(0.01901)	(0.00612)
N	40417	40417	40417
R-Squared	0.1877	0.1872	0.0306
F-statistic	83.143	82.8713	11.3706

Notes — The dependent variable is is the percentage of contracts of that type. The standard errors are robust. All regressions are two-stage least-squares with fixed effects for the agency, good or service and type. The first stage for the TSLS is identical to first-stage shown in the the main table for $P(disp_i = 1|X_i)$.

The Dot-Guessing Game: A "Fruit Fly" for Human Computation Research

John J. Horton
Harvard University
383 Pforzheimer Mail Center
56 Linnaean Street
Cambridge, Massachusetts 02138
john.joseph.horton@gmail.com

ABSTRACT

I propose a human computation task that has a number of properties that make it useful for empirical research. The task itself is simple: subjects are asked to guess the number of dots in an image. The task is useful because: (1) it is simple to generate examples; (2) it is a CAPTCHA; and (3) it has an objective solution that allows us to finely grade performances, yet subjects cannot simply look up the answer. To demonstrate its value, I conducted two experiments using the dot-guessing task. Across both experiments, I found that the "crowd" performed well, with the arithmetic mean guess besting most individual guesses. However, I found that subjects displayed well-known behavioral biases relevant to the design of human computation systems. In the first experiment, subjects reported whether the number of dots in an image was larger or smaller than a randomly chosen number. The randomly chosen reference number strongly anchored subjects' subsequent guesses. In the second experiment, subjects were asked to choose between an incentive payment contract that rewarded good work (and self-knowledge about relative ability) and a fixed payment contract. I found that selection of the incentive contract did not improve performance, nor did it reveal information about relative ability. However, subjects who chose the incentive contract were more likely to be male, suggesting that their choice in contracts was determined by differences in risk-aversion.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics; J.m [**Computer Applications**]: Miscellaneous

General Terms

Design, Economics

Keywords

Crowdsourcing, Experimentation, Mechanical Turk

1. INTRODUCTION

Although new, the field of human computation (HCOMP) has already produced outsized benefits, including spam reduction, digitization of books [17], improved searchability of images [16], and entertainment. Even unintended consequences of this research, such as people solving reCAPTCHAs for pay, while not Pareto-improving, have positive distributional properties—workers in the developing world get cash, people in the developed world are slightly inconvenienced by spam, and books are digitized.[1]

The obvious material benefits of HCOMP alone are enough to stimulate research, but the field could also be a boon to social science. In designing efficient mechanisms, researchers may gain insight into classic questions about motivation and incentives, the effects of peers and norms, and the interplay between organizational structure and outcomes.

Of course, the benefits flow in both directions: psychology, economics, and sociology have much to offer HCOMP. Designers of HCOMP systems must provide quality work, create elicitation procedures, specify the nature of inter-subject interactions (if any), motivate subjects to participate and to reveal their abilities, and decide how to aggregate inputs. Social science may offer useful insights into ways of approaching many of these design challenges.

1.1 Previous work

Several papers have used online labor markets such as Amazon's Mechanical Turk (MTurk) to conduct experiments [9, 15, 14]. Horton, Rand, and Zeckhauser discuss the social science potential of online experiments in these markets, focusing on how challenges to validity can be overcome [6]. There already exists a small literature on crowdsourcing from a social science perspective [7, 11, 5, 3]. New tools are also being developed that make experimentation easier [10].

1.2 Overview

In this paper, I argue that one way to make social science/HCOMP collaboration more fruitful is to identify a "model task" for empirical research, analagous to the "model organisms" in the biological sciences, such as E. coli, the fruit fly, and the nematode worm. Although not considered in-

[1] "Spammers Pay Others to Answer Security Tests," The New York Times, April 25, 2010, http://www.nytimes.com/2010/04/26/technology/26captcha.html.

trinsically interesting, these organisms possess certain properties that make them especially attractive for research purposes. I propose a dot-guessing game, where subjects guess the number of dots in a computer-generated image, as a potential model task. In order to demonstrate its utility and generate new HCOMP-relevant research, I used this task in two experiments conducted on MTurk. The experiments were designed to explore practical HCOMP/crowdsourcing issues: the side-effects of certain forms of elicitation, the potential of using type-revealing contract choices to infer ability and/or (justified) confidence, and the relationships among incentives, effort, and quality.

1.3 Main Experimental Findings

In general, subjects did remarkably well at the task, confirming the "wisdom of crowds." After excluding some egregiously poor guesses, the mean guess was consistently better than most individual guesses. Weighting subjects based on their performance on an initial screening task improved prediction accuracy substantially, though all of the gain came from excluding very poor performers.

In the first experiment, I found that subjects were highly susceptible to "anchoring effects." Before offering a guess, subjects were asked if the true number of dots was above or below some randomly chosen number. This above/below number had a strong effect on subsequent guesses. In the second experiment, I found that offering "high-powered" incentive contracts tying payment to performance were largely ineffective. Uptake of the contract did not reveal which subjects were actually good at the task. Uptake was, however, correlated with gender. Both the anchoring result and the gender/risk preference result have been noted in behavioral economics experiments.

2. A MODEL TASK?

There is great diversity among HCOMP tasks, evidenced by the differences among tasks such as image labeling, transcription, translation, and evaluation/filtering. If we consider HCOMP more broadly, the list grows to include such tasks as offering subjective probabilities (e.g., through prediction markets or scoring rules) and generating new labels, slogans, designs, etc.

It would be convenient to have a single task in experimental studies that could generate results generalizable to other domains. Such a task could improve our ability to compound knowledge and lead to more replications, thereby increasing our confidence in findings. It seems unlikely that a universally appropriate task exists. However, there are certain commonalities among HCOMP tasks and, at least for basic research, certain tasks might partially fill the model organism role.

What are some desirable characteristics for such a task? It should be culturally neutral and easy to explain, even to subjects with poor literacy skills. The quality of worker output should be objective and easy to measure, yet quality should be naturally heterogeneous. "Natural" variation in quality is necessary, not only in order to make it easy to identify treatment effects from different manipulations, but also because one of the key challenges in mechanism design is reliably identifying and overweighting top performers and reducing the influence of poor performers.

Ideally, quality should be a nearly-continuous variable, permitting fine-grained distinctions which make competitive play possible at both the dyad and group levels. Continuous measures minimize the chance of ties, making it easier to offer precise relative-performance contracts or to display subjects' relative positions on a "leader board." Continuous quality measures also simplify analysis by permitting the use of linear models instead of the more complex models needed to analyze dichotomous outcomes. In order to investigate learning, the task should allow subjects to improve with effort or training. There should also be positive relationships between incentives and effort and between effort and quality.

Because so many HCOMP tasks deal with aggregating bits of information, the model task should have a "wisdom of crowds" potential. At the same time, there might also be cases where groupthink, herding, or the "madness of crowds" could prevail. Ideally, subjects should be able to share information and to respond to information about the task provided by others. To measure group-level performance, work outputs should be easy to aggregate so that different aggregation approaches can be tested.

Finally, logistically, we want to be able to generate new, unique tasks with ease. Any task should—like a CAPTCHA—be a hard AI problem, especially if we try to use high-powered incentives that might encourage cheating via algorithms.

2.1 Dot-guessing game

In one of the first "wisdom of crowds" demonstrations, Francis Galton analyzed the results of a contest where fairgoers guessed the weight of a butchered ox. While most individual guesses were far off the mark, the mean guess was remarkably accurate. It is easy to generate a 21st-century version of this game for play over the Internet: subjects are asked to guess the number of dots in a computer-generated image, an example of which is shown in Figure 1 (along with the four lines of R code needed to generate the image). Creating new and unique tasks in any computer language is only marginally more challenging than the "hello world" program.

Aside from ease of creation, the dot-guessing game has a number of desirable properties as a research tool. It is very simple to explain. It has an unambiguous quality metric—how far off a guess is from the correct answer—and yet answers cannot simply be looked up online or solved (easily) with a computer. It is in some ways similar to the "pseudo-subjective" questions psychologists have used to measure over-confidence, such as asking subjects to estimate the length of the Nile river or the gestation period (in days) of an Asian elephant [2].

Obviously the dot-guessing game also has limitations. It is not good for generative tasks (e.g., image labeling). Nor is it ideal for studying coordination, where tasks like the graph-coloring problem may be more appropriate [8]. However, for a wide range of human computation scenarios that require judgment by workers and the screening and aggregation of inputs, the dot-guessing game offers many advantages.

```
> par(mar = c(0, 0, 0, 0))
> n = 500
> X = c(runif(n, 0, 1), runif(n, 0, 1))
> plot(X, axes = F, xlab = "", ylab = "")
```

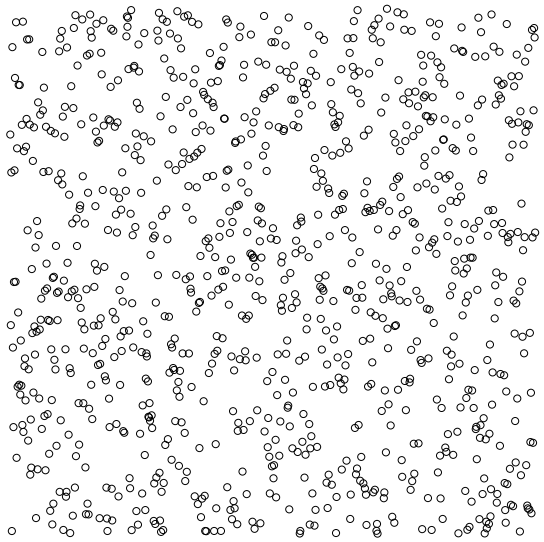


Figure 1: Example dot-guessing game with generating R code.

3. EXPERIMENT A

In the first experiment, 200 subjects were recruited to complete a single HIT. They each inspected a single image containing 500 dots and offered a guess. Before proffering a guess, each subject first answered whether they thought the number of dots was greater than or less than a number X randomly drawn from a uniform distribution, $X \sim U[100, 1200]$.

3.1 Results

Overall, subjects performed quite well, with the mean besting most individual guesses once very bad outliers were removed. Table 1 shows the performance results for the"Anchor Experiment" (as well as for the follow-on experiment) after removing subjects with an error *rate* of 10 or greater. The experiment highlights the dangers of free-response elicitation: one subject estimated that the photograph contained 15823 dots—an over-estimate by a factor of over 31.

Table 1: Dot-guessing performance

Dots		Mean	Geo. Mean	Med.	SD	N	Quant.
Anchor Experiment							
500		517	435	450	315	195	0.93
Incentive Experiment							
310	(1)	375	285	292	349	158	0.69
456	(2)	537	373	400	554	148	0.74
854	(3)	986	579	606	1454	138	0.83
1243	(4)	1232	866	850	1443	143	0.99
2543	(5)	2362	1684	1700	2580	147	0.94
4000	(6)	4888	2855	2562	5970	144	0.92
8200	(7)	8559	4748	4644	12267	134	0.96

Notes: Data excludes subjects with error rates greater 10. "Quant." reports the location of the error associated with the mean guess in the distribution of errors associated with all guesses.

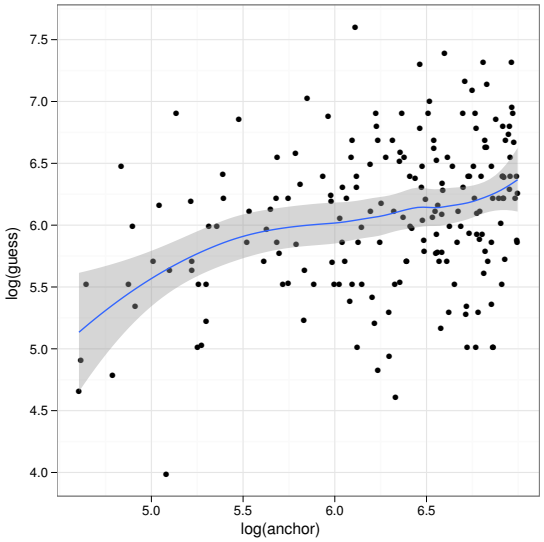


Figure 2: Log guesses versus log anchors. Subjects' guesses for a 500-dot image are plotted against their anchors.

3.1.1 *Anchoring effects*

Despite the good aggregate performance, subjects' guesses about the 500-dot image were strongly affected by asking whether the number of dots was above or below the randomly chosen anchors: a 10-point increase in the anchor increases guesses by approximately 3 points. Regressing guesses on the anchors gives

$$y_i = \underbrace{0.283}_{0.08} \cdot X_i + \underbrace{341.841}_{52.38}$$

with $R^2 = 0.07$ and $N = 194$. Standard errors are shown underneath the coefficient estimates. Figure 2 illustrates the guess-anchor relationship through a scatter plot of the logged guesses versus the logged anchors and a local kernel regression line.

4. EXPERIMENT B

The second experiment was designed to test whether incentive contracts reveal information about subject confidence and/or motivate better performance. Subjects were simultaneously asked to guess the number of dots in an image and to state their preference between two incentive contracts: (1) \$5.00 if the subject's answer was in the top half of all guesses, \$0 otherwise; or (2) \$2.50 for certain. The response was coded as $risk_{ij} = 1$ if the subject chose the incentive contract, $risk_{ij} = 0$ if the subject chose the fixed contract. Similarly, if a subject i was in the top half of the distribution for image j, then $top_{ij} = 1$. Subjects were allowed to

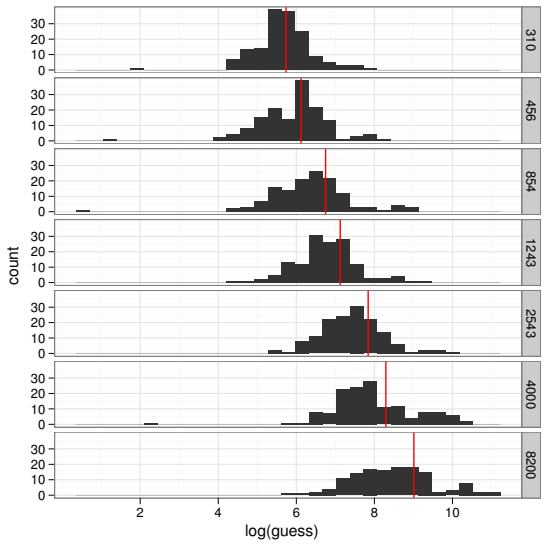


Figure 3: Distributions of log guesses with correct answers indicated by vertical bars.

view up to seven images, each of which contained a different number of dots. Having multiple responses per subject makes a multilevel model appropriate for analysis [4], and all the regressions include individual-specific random effects.

Subjects were not truly randomly assigned to HITs, but rather were randomly assigned a HIT from the pool of uncompleted HITs. When a worker "returned" a task (i.e., decided not to complete it), it was returned to the pool. Over time, the portion of images with high numbers of dots in the pool increased, as workers were more likely to return those images. The reasons workers found those images more burdensome are unclear. The most likely explanation is that the images simply took longer to download, due to larger file sizes: in a short follow-on experiment using images of equal file size containing different numbers of dots, differential attrition disappeared.

To minimize the burden of answering survey questions, subjects were randomly assigned to be asked one of the following demographic questions or no question at all: (1) age, (2) gender, or (3) both age and gender.

4.1 Results

Figure 3 shows the histograms of log guesses for each of the seven images. The logs of the correct answers are shown by vertical red lines. In general, guesses appear to be log normally distributed. Among the images with relatively fewer dots, the distributions appear to be symmetric around the correct answers. The symmetry breaks down among the images with larger numbers of dots, and there is evidence of systematic under-guessing combined with a few very large positive outliers.

4.1.1 Contract Choice

Few subjects selected to take the incentive contract, with mean uptake only 22.43%. Regressing contract choice on an indicator for being in the top of the distribution, we have

$$risk_{ij} = \underbrace{-0.00}_{0.02} \cdot top_{ij} + \underbrace{0.25}_{0.03} + \dots$$

with $R^2 = 0.74$ and $N = 1012$. As with all the regressions examining the second experiment, this regression includes picture fixed effects and individual random effects. We can see that the choice of contract does not appear to reveal anything meaningful about the quality of the subjects' responses. Unsurprisingly, the error rate is also unrelated to contract choice. Regressing E_{ij} on the incentive contract indicator and dummy variables for each of the seven images (coefficeints are omitted below), there is no apparent relationship

$$E_{ij} = \underbrace{0.03}_{0.09} \cdot risk_{ij} + \dots$$

with $R^2 = 0.65$ and $N = 1012$.

As we can see from the regression line, workers who chose the incentive contract performed slightly worse, but the effect was not significant. However, those workers did spend more time on the task, though the effect was not significant at conventional levels:

$$\log time_{ij} = \underbrace{0.10}_{0.06} \cdot risk_{ij} + \dots$$

with image-specific fixed effects and individual level random effects, giving $R^2 = 0.98$ and $N = 1012$. Note that the high R^2 in this regression results from the inclusion of the fixed effects of the image. Because the task itself took so little time, most of the variation in times is likely the result of differences in download time. Although in this experiment there was no relationship between time spent and work quality, this is certainly not a general result. Adding an incentive contract could be beneficial, even if the choice of contract is not type-revealing.

4.2 Gender and contract choice

Regressing contract choice on gender, and including image-specific fixed-effects and individual-specific random effects, we have

$$risk_{ij} = \underbrace{0.11}_{0.05} \cdot male_{ij} + \dots$$

with $R^2 = 0.75$ and $N = 949$. Although both male and female subjects preferred the fixed contract, females had a stronger preference for the risk-free contract. This is consistent with experimental evidence that females are more risk-averse when playing abstract games.[2] This finding of gender difference is not in itself surprising, but it does suggest that using incentive contract uptake to sort individuals is problematic and could lead to inefficient, gender-based discrimination.

In fact, risk-seeking male subjects actually performed slightly

[2]However, others have argued that this gender difference disappears when decisions are contextualized as real business decisions about things like investments and insurance [13].

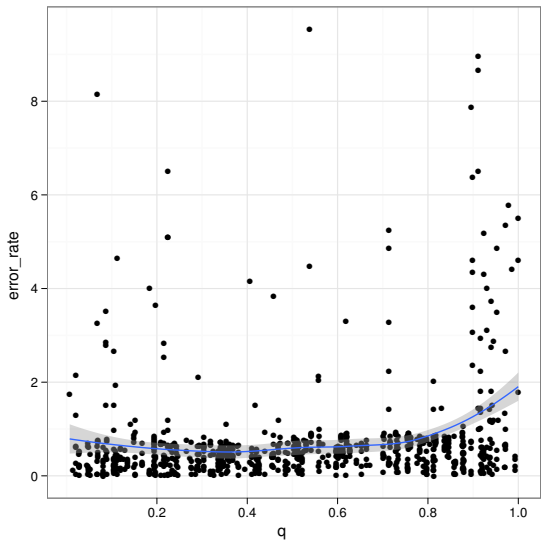


Figure 4: Relative performances (percentiles) on initial tasks versus error rates on subsequent tasks.

worse than females, although the effect is not significant

$$E_{ij} = \underbrace{0.09}_{0.13} \cdot male_{ij} + \dots$$

with image fixed effects and individual random effects, giving $R^2 = 0.66$ and $N = 949$.

5. IMPROVING ESTIMATES

To improve accuracy in applications, one would presumably try to give more weight to the output of better workers. To demonstrate how this can be done with the dot-guessing task, I weighted subjects based on relative performance on the first task completed. For each subject who completed more than one image, I computed their error on the first image and where the performance placed them in the distribution of all performances for that image.

Figure 4 is a scatter plot of subjects' relative performances on their respective initial tasks versus error rates for subsequent tasks. A local kernel density estimate shows that subsequent performance is fairly uniform up to about the 90th percentile, perhaps with some evidence of mean reversion for very low quantiles, and that after the 90th percentile performance is much worse, with many subjects achieving error rates of 200% or more.

To improve performance, I used a very simple weighting method with just two parameters. I split the distribution of errors at s, and then gave all subjects who performed better than s a weight of 1, and all who performed worse than s a weight of h, with $h < 1$. Using a grid search, I found that the mean error is minimized when $s = .95$ and $h = 0$. Because the quantiles (and, hence, the parameters) were determined using a subject's first image, using this weighting scheme for all observations would lead to a mechanical reduction in the error rate. Instead, I computed the mean guess (weighted and unweighted) without using any first responses. Table 2 shows that, in general, adjusting the estimate leads to improvement in accuracy. In six out of seven cases, the mean error of the adjusted estimate was lower than the unadjusted estimate.

Table 2: Effects of ex post weighting

Image	Dots	Unadj.	Adj.	Chg.
1	310	344 (0.11)	318 (0.027)	+
2	456	529 (0.16)	485 (0.063)	+
3	854	941 (0.102)	855 (0.002)	+
4	1243	1267 (0.019)	1252 (0.007)	+
5	2543	2366 (0.07)	2345 (0.078)	-
6	4000	4839 (0.21)	4648 (0.162)	+
7	8200	9683 (0.181)	9473 (0.155)	+

Notes: Estimates using optimal weighting function ($s = .95, h = 0$). The error rate, $|est. - truth|/truth$, is in parentheses. Estimates do not include the subjects' first tasks completed.

6. CONCLUSION

The results generally confirm the "wisdom of crowds" hypothesis. By excluding the worst performers, the mean guesses were quite accurate, though performance worsened as the number of dots in the image increased.

In terms of the treatment effects, subjects were prone to the anchoring bias: when subjects were asked whether the number of dots in an image was above or below some randomly chosen number, the number strongly affected subjects' guesses.

I found that subjects to do not reveal their type (i.e., relative aptitude for the task) through their preference in contract type. Males were more likely to choose an incentive contract, but they performed no better than females. There is some evidence that subjects who chose the incentive contract spent more time on the tasks, though it is unclear which way causality ran. Ex post processing of the data improved prediction accuracy: removing the worst 5% of subjects led to improvements in accuracy, reducing the error in six of the seven cases (and essentially equalizing the results of the seventh case).

6.1 Implications

The results suggest that mechanism designers should consider that: (a) choice of contract type may not be informative, and (b) differences in risk appetite (which may be systematic and/or have a gender component) may lead to mechanisms that inefficiently overweight the over-confident. Because I found a positive correlation between time spent and preference for an incentive contract, it may be the case that incentive contracts improve quality, though contract choice was endogenous, as was the time spent on the task.

The anchoring results suggest that mechanism designers must be cautious in querying subjects, even in a hypothetical manner. It seems possible that subjects will accept those questions as informative and adjust their beliefs rather than treat them as irrelevant (in the sense of not providing new

information).

6.2 Case for Empiricism

As more aspects of social life moves online, both the size of collected observational data and the potential to conduct experiments will continue to grow. The transition to a bench science, where hypotheses can be quickly tested through experimentation, is bound to have positive effects.

The case for empiricism is particularly strong in designing mechanisms, as theory offers little guidance about cognitive demands on the mechanism, how contextual factors are likely to influence subjects, how boundedly rational workers actually play, etc. To make practically useful mechanisms, it is important to understand how they hold up when the players are real people, prone to the biases and subject to the limitations of real people [12].

7. ACKNOWLEDGMENTS

Thanks to the the Christakis Lab at Harvard Medical School and the NSF-IGERT Multidisciplinary Program in Inequality & Social Policy for generous financial support (Grant No. 0333403). Thanks to Robin Yerkes Horton, Brendan O'Connor, Carolyn Yerkes, and Richard Zeckhauser for very helpful comments. All plots were made using `ggplot2` [18] and all multilevel models were fit using `lme4`[1].

8. REFERENCES

[1] D. Bates and D. Sarkar. lme4: Linear mixed-effects models using S4 classes. *URL http://CRAN. R-project. org/package= lme4, R package version 0.999375-28*, 2008.

[2] B. Biais, D. Hilton, K. Mazurier, and S. Pouget. Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market. *Review of Economic Studies*, pages 287–312, 2005.

[3] D. L. Chen and J. J. Horton. The wages of paycuts: Evidence from a field experiment. *Working Paper*, 2009.

[4] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press Cambridge, 2007.

[5] J. J. Horton and L. B. Chilton. The labor economics of paid crowdsourcing. *Proceedings of the 11th ACM Conference on Electronic Commerce 2010 (forthcoming)*, 2010.

[6] J. J. Horton, D. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *SSRN eLibrary*, 2010.

[7] B. A. Huberman, D. Romero, and F. Wu. Crowdsourcing, attention and productivity. *Journal of Information Science (in press)*, 2009.

[8] M. Kearns, S. Suri, and N. Montfort. An experimental study of the coloring problem on human subject networks. *Science*, 313(5788):824, 2006.

[9] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. *Proceedings of Computer Human Interaction (CHI-2008)*, 2008.

[10] G. Little, L. B. Chilton, R. Miller, and M. Goldman. Turkit: Tools for iterative tasks on mechanical turk. *Working Paper, MIT*, 2009.

[11] W. Mason and D. J. Watts. Financial incentives and the 'performance of crowds'. In *Proc. ACM SIGKDD Workshop on Human Computation (HCOMP)*, 2009.

[12] D. McFadden. The human side of mechanism design: A tribute to Leo Hurwicz and Jean-Jacque Laffont. *Review of Economic Design*, 13(1):77–100, 2009.

[13] R. Schubert, M. Brown, M. Gysler, and H. W. Brachinger. Financial decision-making: Are women really more risk-averse? *The American Economic Review*, 89(2):381–385, 1999.

[14] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label?: Improving data quality and data mining using multiple, noisy labelers. *Knowledge Discovery and Data Minding 2008 (KDD-2008)*, 2008.

[15] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, 2008.

[16] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

[17] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321(5895):1465, 2008.

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

The Need for Standardization in Crowdsourcing

Panagiotis G. Ipeirotis
New York University
panos@stern.nyu.edu

John J. Horton
Harvard University
oDesk Corporation
john.joseph.horton@gmail.com

ABSTRACT

Crowdsourcing has shown itself to be well-suited for the accomplishment of certain kinds of small tasks, yet many crowdsourceable tasks still require extensive structuring and managerial effort before using a crowd is feasible. We argue that this overhead could be substantially reduced via standardization. In the same way that task standardization enabled the mass production of physical goods, standardization of basic "building block" tasks would make crowdsourcing more scalable. Standardization would make it easier to set prices, spread best practices, build meaningful reputation systems and track quality. All of this would increase the demand for paid crowdsourcing—a development we argue is positive on both efficiency and welfare grounds. Standardization would also allow more complex processes to be built out of simpler tasks while still being able to predict quality, cost and time to completion. Realizing this vision will require interdisciplinary research effort as well as buy-in from online labor platforms.

Bios

Panos Ipeirotis has been using Mechanical Turk for his research (mainly to gather relevance judgments) since 2006, and over time the platform transformed from a research tool into a topic for research. He is interested in quality management and scalability issues related to crowdsourcing.

John Horton uses economics to explore the growing phenomena of people working online. He is particularly interested in the use of online labor markets as tools for economic development.

INTRODUCTION

The academic community and a growing number of firms are looking to paid crowdsourcing to solve problems. The problems being solved vary, but what they all have in common is one or more sub-problems that cannot be fully automated, and require human labor. This labor demand is being met by workers recruited from online labor markets such as Amazon Mechanical Turk, Microtask, oDesk and Elance or from casual participants recruited by intermediaries like CrowdFlower and CloudCrowd. In these markets, buyers and sellers have great flexibility in the tasks they propose and the making and accepting of offers.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

The flexibility of online labor markets is similar to the flexibility of traditional labor markets. In both markets, buyers and sellers are free to trade almost any kind of labor at almost any terms. However, an important distinction between online and offline is that once a worker is hired off an offline, traditional market, they are *not* allocated to tasks via a spot market. Workers within firms are employees who have been screened, trained for their jobs and are have incentives for good performance—at a minimum, poor performance can cause them to lose their jobs. Furthermore, for many jobs—particularly those focusing on the production of physical goods—good performance is very well defined, in that workers must adhere to a standard set of instructions. This standardization of tasks is the essential feature of modern production and how it can be applied to crowdsourcing is the focus of our paper.

With task standardization, innovators like Henry Ford could ensure that hired workers—after suitable training—could complete those tasks easily, predictably and in a way that workers can be easily replaced with others, similarly trained. To return to paid crowdsourcing, most of the high demand crowdsourcing tasks are low-skilled and require workers to closely and consistently adhere to instructions for a particular, standardized task. As it currently stands, existing crowdsourcing platforms bear little resemblance to Henry Ford's car plants. In crowdsourcing markets, the factory would be more like an open bazaar where workers could come and go as they pleased, receiving or making offers on tasks that different in their difficulty and skill requirements ("install engines!", "add windshields!", "design a new chassis!") for different rates of pay—and with different pricing structures (fixed payment, hourly wages, incentives etc.). Some buyers would be offering work on buses, some on cars, some on lawnmowers. Reputations would be weak and easily subverted. Among both buyers and sellers, one can find scammers; some buyers are simply recruiting accomplices for nefarious activities.

The upside of such a disorganized market is that workers and buyers have lots of flexibility. There are good reasons for not wanting to just recreate the on-line equivalent of single-firm factory. However, we do not think it is an "either-or" proposition. In this paper, we discuss ways that we can have more structure on a marketplace platform, without undermining its key advantages. In particular, we believe that greater task standardization, a cultivated garden approach to work-pools and a market-making type work allocation mechanism to help arrive at prices could help us build scalable human-powered systems that meet real-world needs.

The rest of the paper is structured as follows. First, we give a

brief overview of the current status of on-line labor marketplaces. Next, we argue that the standardization of simple tasks can lead to immediate benefits in terms of pricing, speed, and easy adoption of best practices. Then, we discuss what are the benefits of constructing advanced tasks using workflows of standardized tasks, and discuss the role of the marketplace owners in properly designing the market for optimal outcomes for all participants. Finally, we conclude by presenting our views on the nature of platforms and other third parties (including researchers) and sketch out areas for future research.

CURRENT STATUS

In addition to the academic interest in paid crowdsourcing, an expanding ecosystem of firms use crowdsourcing for some business task, offer themselves as a platform for labor or help intermediate between labor pool and would-be users of crowdsourcing. A crowdsourcing industry group, Crowdsortium was recently formed and some of the larger players recently organized a conference, CrowdConf that was well-attended, attracting start-ups, academics and investors.

Despite the excitement and apparent industry maturation, there has been relatively little innovation—at least at the micro-work level—in the technology of how workers are allocated tasks, how reputation is managed and how tasks are presented etc. As innovative as MTurk is, it is basically unchanged since its launch. The criticism of MTurk—the difficulty of pricing work, the difficulty in predicting completion times and gaining quality, the inadequacy of the way that workers can search for tasks—are recurrent and still unanswered. Would-be users of crowdsourcing often fumble, with even technically savvy users getting mixed results. Best practices feel more like folk wisdom than an emerging consensus. Even more troubling, there is some evidence that at least some markets are becoming inundated with spammers.

One part of the crowdsourcing ecosystem that appears to be thriving is the "curated garden" approach used by companies like uTest (testing software), MicroTask (quality assurance for data entry), CloudCrowd (proofreading and translation), and LiveOps (call centers). These firms recruit and train workers for their standardized tasks and they set prices of both sides of the market. Because the task is relatively narrow, it is easier to build meaningful, informative feedback and verify ex ante that workers can do the task, rather than try to screen bad work out ex post. While this kind of control is not free, practitioners gain the scalability and cost savings of crowdsourcing without the confusion of the open market. The downside of these curated pools is that access as both a buyer and seller is limited. One of the great virtues of more market like platforms is that they are democratic and easy to experiment on. The natural question is whether it is possible to create labor pools that look more like curated gardens—with well defined, standardized tasks—and yet are still relatively open, both to new buyers and sellers?

STANDARDIZING BASIC WORK UNITS

Currently, the labor markets operate in a completely uncoordinated manner. Every employer generates its own work request, prices the request independently, and evaluates the answers separately from everyone else. Although this approach have some intuitive appeal in terms of worker and employer flexibility, it is a fundamentally inefficient approach.

- Every employer has to implement from scratch the "best practices" for each type of work. For example, there are multiple UI's for labeling images, or for transcribing audio. The long-term employers learn from their mistakes and fix the design problems, while newcomers have to learn the lessons of bad design the hard way.
- Every employer needs to price its work unit without knowing the conditions of the market and this price cannot fluctuate without removing and reposting the tasks.
- Workers need to learn the intricacies of the interface for each separate employer.
- Workers need to adapt to the different quality requirements of each employer.

We believe that the efficiency of the market can increase tremendously if there is at least some basic standardization of the common types of (micro-)work that is being posted on online labor markets.

So, what are these common types of (micro-)work that we can standardize? Amazon Mechanical Turk lists a set of basic templates[1], which give a good idea of what tasks are good candidates to standardize first. The analysis of the Mechanical Turk marketplace [4] also indicates a set of tasks that are very frequent on Mechanical Turk and are also good candidates to standardize.

We can draw in parallel with engineering: In mechanics, we have a set of "simple machines"[2], such as screws, levers, wheel and axle, and so on. These simple machines are typically standardized and serve as components for larger, significantly more complicated creations. Analogously, in crowdsourcing, we can define a set of such simple tasks, standardize them, and then build, if necessary, more complicated tasks on top.

What are the advantages of standardizing the simple tasks, if we only need them as components?

First of all, as mentioned above, there is no need for requesters to think on how to create the user interfaces and best practices for such simple tasks. These standardized tasks can be, of course, revised over time to reflect our knowledge on how to best accomplish them.

Second, and potentially more important, these simple tasks can be traded in the market in the same way that stocks and commodities are currently traded in financial markets. In stock markets, the buyer does not need to know who is the seller, or whether the order was fulfilled by a single seller or multiple ones: it is the task of the market maker to match and fulfill buy and sell orders. In the same way, we can have a queue of standardized tasks that need to be completed, and workers can complete them at any time, without having to think about the reputation of the requester or to (re-)familiarize themselves with the task. This should lead to much more efficient task execution.

A third advantage of standardized work units is that pricing becomes significantly simpler. Instead of "testing the market" to

[1] https://requester.mturk.com/bulk/hit_templates
[2] http://en.wikipedia.org/wiki/Simple_machine

see what price points leads to an optimal setting, we can instead have a very "liquid" market with a large number of offered tasks and a large number of workers that work on these tasks. This can lead to a stock-market-like pricing. The tasks get completed by the workers, in priority order according to the offered price for the work unit: the highest paying units get completed first. So, if requesters want to prioritize their own tasks, they can simply price them higher than the current market price. This corresponds to an increase in demand, which moves up the market price. On the other hand, if no requesters post tasks then, once the tasks with the highest prices get completed, then we automatically move to the tasks that have lower price associated with them. This corresponds to the case where the supply of work is higher than the demand, and market prices for the work unit move down.

In cases where there is not enough "liquidity" in the market (i.e., when the workers are not willing to work for the posted prices), then we can employ automated market makers [6], such as the ones currently used by prediction markets. The process would then operate like this: The workers identify the price for which they are willing to work. Then, the automated market maker takes into consideration the "ask" (the worker quote) and the "bid" (the price of the task), and can perform the trade by "bridging" the difference. Essentially, such automated market makers provide a subsidy in order for the transactions to happen. We should note that a market owner can typically benefit even in scenarios, where they need to subsidize the market through an automated market maker: the fee from a transaction that happens can cover the necessary subsidy which is consumed by the automated market maker.

We believe that having basic, standardized work units with highly liquid, high-volume markets can serve as a catalyst for companies to adopt crowdsourcing. Standardization can strengthen the network effects, can provide the basis for better reputation systems, can facilitate pricing, and can lead to the easier development of more complicated tasks that comprise of an arbitrary combination of small work units. We describe how to leverage basic work units next.

CONSTRUCTING AND PRICING COMPOSITE TASKS

Once we have some basic work units in place, we can start generating tasks that consist of multiple such units, to generate tasks that cannot be achieved with just using basic units.

Again we can draw the analogs from mechanical engineering: the "simple machines" (screws, levers, wheel and axle, and so on) can then be assembled together to generate machines of arbitrary complexity. Similarly, in crowdsourcing we can use these *standardized* set of "simple work units" that can be later assembled to generate tasks of arbitrary complexity.

Quality Assurance: Assume that we have a basic work unit for a task such as comment moderation, that guarantees an accuracy of 80% or higher (e.g., by screening and testing continuously the workers that can complete these tasks). If we want to have a work unit that has higher quality guarantees, we can generate a composite unit that uses multiple, redundant work units and relies on, say, majority vote to generate a work unit with higher quality guarantees.

Pricing Workflows: There is already work available on how to create [5] and control the quality [2] of workflows in crowdsourced environments. We also have a set of design patterns for workflows in general.[3] If we have a crowdsourced workflow that consists of standardized work units, we can also accurately price the overall workflow. We do not even have to reinvent the wheel: there is a significant amount of work on pricing combinatorial contracts [1] in prediction markets.[4] A workflow can be expressed as a combinatorial expression of the underlying simple work units. Since we know the price of standard units, we can easily leverage work from prediction markets to price tasks of almost arbitrary complexity. The successful deployment of Predictalot[5] by Yahoo! during the 2010 soccer World Cup, with the extensive real-time pricing of complicated combinatorial contracts, gives us the confidence that such a pricing mechanism is also possible for online labor markets.

Timing and Optimizing Workflows: There is already significant amount of work in distributed computing on optimizing execution of task workflows in Mapreduce-like environments [3]. This research should be directly applicable in an environment where the basic computation is performed not by computers but by humans. Also, since the work units will be completed through easy-to-model waiting queues, we can easily leverage the work from queuing theory to estimate how long a task will remain within the system: by identifying the critical parts of execution we can also identify potential bottlenecks and increase the offered prices for only the work units that critically affect the completion time of the overall task.

ROLE OF PLATFORMS

One helpful way to think about the role and incentives of on-line labor platforms is to consider that they are analogous to a commerce-promoting government in a traditional labor market. Most platforms levy an ad valorem charge and thus they have an incentive to increase the size of the total wage bill. While there are many steps these markets can take, their efforts fall into two categories: (1) remedying externalities and (2) setting enforceable standards and rules, i.e., their "weights and measures" function.

Remedying Externalities

An externality is created whenever the costs and benefits from some activity are not solely internalized by the person choosing to engage in that activity. A negative example is pollution—the factory owner gets the goods, others get the smoke—while a positive example is yard beautification (the gardener works and buys the plants, others get to enjoy the scenery). Because the parties making the decision do not fully internalize the costs and benefits, activities producing negative externalities are (inefficiently) over-provided, and activities producing positive externalities are (inefficiently) under-provided. In such cases, "government" intervention can improve efficiency.

Negative examples are easy to find in on-line labor markets—fraud is one example. Not only is fraud unjust, it also makes

[3] http://www.workflowpatterns.com/patterns/

[4] An example of a combinatorial contract: "Obama will win the 2012 election *and* will win Ohio" or "Obama will win the 2012 election *given that* he will win Ohio"

[5] http://predictalot.yahoo.com/

everyone else more distrustful, lowering the volume and value of trade. Removing bad actors helps ameliorate the market-killing problem of information asymmetry, as uncertainty about the quality of some good or service is often just the probability that the other trading partner is a fraud.

A positive example is honest feedback after a trade. Giving feedback is costly to both buyers and sellers: It takes time and giving negative feedback invites retaliation or scares off future trading partners. In the negative case, the platform needs to fight fraud—not simply fraud directed at itself but fraud directed at others on the platform, which has a negative second-order effect on the platform creator. In the positive case, the firm can make offering feedback more attractive, by offering rewards, making in mandatory, making it easier, changing rules to prevent retaliation etc.

There are lots of options in both the positive and negative case—the important point is that platform creators recognize externalities and act to encourage positive externalities and eliminate the negative ones. Individual participants do not have the incentives (or even the ability) to fix the negative externalities for all other market participants. For example, no employer has the incentive to publish his own evaluation of the workers that work for his, as this is a signal earned after a significant cost for the employer. This is a case where the market owner can provide the appropriate incentives and designs for the necessary transparency.

Setting Enforceable Standards

Task standardization will probably require buy-in from on online labor markets and intermediaries. Setting cross-platform standards is likely to be a contentious process, as the introduction of standards gives different incentives to different firms, depending upon their business model and market share. However, at least within a particular platform and ignoring their competitors, there is powerful incentive to create standards as they raise the value of paid crowdsourcing and promote efficiency. For example, the market for SMS's took off in the US only when the big carriers agreed on a common interoperable standard for sending and receiving SMS's across carrier's networks.

In traditional markets, market-wide agreement about basic units of measure facilitate trade. In commodity markets, agreements about quality standards serve a similar role, in that buyers know what they are getting and sellers know what they are supposed to provide. (For example, electricity producers are required to produce electricity adhering to some minimum standards before being able to connect to the grid and sell to other parties.) It should be clear that having public standards make quality assurance easier for the platform: enforcing standards on standardized units of work can be done much easier than enforcing quality standards in a wide variety of adhoc tasks. With such standards, it easier to imagine platform owners more willingly taking the role of testing for and enforcing quality standards for the participants that provide labor.

If we define weights and measures more broadly to include verification of claims, the platform role becomes even wider. They can verify credentials, test scores, work and payment histories, reputation scores and every other piece of information that individuals cannot credibly report themselves. Individuals are also not able to credibly report the quality of their work, but at least with an objective standard, validating those claims is possible.[6]

CONCLUSION

As we laid out in the paper, we believe that the key research challenge will be standardizing tasks. As standardization gains ground and best practices emerge, research will shift towards building labor-appropriate reputation systems and auctions as well as more complex work flows. Much of this research can be done at a micro level through researcher-initiated experiments. However, some "macro" research about platform-wide policies and institutions that cannot be tested anywhere but at the platform level. For these kinds of questions, we will need to rely on observational data or on large experiments conducted in conjunction with platform creators. Examples of this kind of research include dispute resolution policies, policies related to search and matching and reputation systems.

As our knowledge increases and platforms and practices mature, we expect far more work to be outsourced to remote workers. On the whole, we think this is a positive development, particularly because paid crowdsourcing gives people in poor countries access to buyers in rich countries, enabling a kind of virtual migration. If this form of increased virtual labor mobility has effects similar to those of increased real labor mobility, then the emergence of online labor markets could be transformative, as the welfare gains from liberalizing restrictions on labor mobility are truly enormous. Because of these potential welfare impacts, we view research on paid crowdsourcing as not only intellectually interesting, but also a way to expand economic opportunities for people who have relatively few opportunities due the accidents of birth and national borders.

References

[1] Y. Chen, S. Goel, and D. Pennock. Pricing combinatorial markets for tournaments. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 305–314. ACM, 2008.

[2] P. Dai, . Mausam, and D. Weld. Decision-theoretic control of crowd-sourced workflows. 2010.

[3] A. F. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava. Building a high-level dataflow system on top of map-reduce: the pig experience. *Proc. VLDB Endow.*, 2:1414–1425, August 2009.

[4] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS*, 17:16–21, December 2010.

[5] G. Little, L. Chilton, M. Goldman, and R. Miller. Turkit: Tools for iterative tasks on mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 29–30. ACM, 2009.

[6] A. Othman, T. Sandholm, D. Pennock, and D. Reeves. A practical liquidity-sensitive automated market maker. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 377–386. ACM, 2010.

[6] For example, one of the main innovations made by oDesk was that they logged a worker's time spent on a task, enabling truthful hourly billing.

Heads in the Cloud

A professor and several PhD students at MIT examine the challenges and opportunities in human computation.

By Robert C. Miller, Greg Little, Michael Bernstein, Jeffrey P. Bigham, Lydia B. Chilton, Max Goldman, John J. Horton, and Rajeev Nayak

DOI: 10.1145/1869086.1869095

Crowd computing is quickly becoming an essential part of the technology landscape. Crowd computing encompasses the interaction among large numbers of people facilitated by software systems and networking technology. Crowds—and by "crowds," we literally mean a mass of people—are themselves the power that fuels sites like Wikipedia, Twitter, Intrade, and even online labor markets like Amazon Mechanical Turk.

One way to think about crowd computing is as the human analogue to cloud computing. Where the cloud provides access to elastic, highly available computation, and storage resources in the network, the crowd represents access to elastic, highly-available *human* resources, such as human perception and intelligence. Crowd computing offers the strength of software with the intelligence and common sense of human beings.

HUMAN COMPUTATION

One variant of crowd computing is human computation, which we define as using software to orchestrate a process of small contributions from a crowd to solve a problem that can't be solved by software alone.

Human computation was first popularized by Games With a Purpose (http://gwap.com), in which the computation is a side effect of a fun game **[8]**. For example, the *ESP Game* asks two players to guess words associated with an image, scoring points when their words agree, which makes the game fun, but also generating useful labels to index the image for searching, which makes it an act of human computation.

Another human computation site is Amazon Mechanical Turk (http://mturk.com), a marketplace where people get paid to perform human computation. Users, or "workers," find short tasks that are posted by "requesters" (the people who need the tasks completed) and get paid small amounts of money for completing them. CrowdFlower (http://crowdflower.com) is another site that pays users for computation—in not only real currency, but also virtual currencies for games like *Farmville* and *Mafia Wars*. Social networks like Facebook are also becoming platforms for human computation, motivated by social relationships rather than entertainment or monetary reward.

These platforms make it increasingly feasible to build and deploy systems that use human intelligence as an integral component. But there are at least three challenges to exploring the space of human computation systems: 1) applications—understanding what's appropriate for human computation and what isn't; 2) programming—learning how to write software that uses human computation; and 3) systems—learning how to get good performance out of a system with humans in the loop.

APPLICATIONS

What application areas will benefit the most from human computation? What properties do certain problems possess that make them amenable to a successful solution by a hybrid human-software system? Since the end user of such a system is also, typically, human, we can refine this question further: Why does a human end user need to request the help of a human crowd to accomplish a goal, rather than just doing it herself?

One reason is differences in capability. A group of many people has abilities and knowledge that one single end user does not, either innately or because of situational constraints. For example, VizWiz **[1]** helps blind us-

ers answer questions they have about things around them that they cannot see. The blind person takes a photograph with a smartphone's camera, records a spoken question (also using the phone), and then uploads the query and picture to a crowd of sighted users on the net who are better able to answer it (see **Figure 1**). For example, if a blind person grabs a can out of her cupboard but has forgotten what's inside it, she can snap a photo of the can and its label, upload it, and ask the sighted users what's in the can.

A related system, Sinch [7], draws on the crowd to provide assistance to web-enabled mobile device users who have situational disabilities, such as the limited ability to read a small screen, arthritis or hand tremors that make it difficult to click on small web page targets, and slow networks. With Sinch, the mobile users speak a question into their phone and the crowd searches the web for answers, using their more capable desktop web access, and returning web pages with the requested information highlighted.

Another reason to use a crowd is the "many eyes" principle, which has been claimed as an advantage of open-source software development (the complete phrase is "many eyes make bugs shallow"). We have exploited this principle in Soylent [2], a Microsoft Word extension that uses a crowd for proofreading, shortening, and repetitive editing. A typical run of Soylent may have dozens of people looking at each paragraph of a document, finding errors that a single writer might miss. In fact, a conference paper submitted about Soylent contained a grammatical error that was overlooked by not only Word's built-in grammar checker, but also eight authors and six reviewers. However, when we passed the paper through Soylent, the crowd caught the error.

A corollary of the many eyes principle is diversity. The fact is, a crowd comprises a wide range of ideas, opinions, and skills. For example in Soylent, the system not only indentifies writing errors, but also suggests multiple ways to fix them. It can suggest text to cut to save space—a tough task even for skilled authors, who are often reluctant to make cuts. Soylent can typically trim text down to 85 percent of its original length, without changing the meaning of the text or introducing errors (see **Figure 2**).

Figure 1: With VizWiz, blind people take photos using their mobile phones and submit them alongside a question, spoken orally into the phone, shown here above each image. A crowd of anonymous users reply, shown below, with response time given in seconds in parentheses.

PROGRAMMING

Prototyping a human computation system is hard if you have to entice a crowd to visit your website. Games With a Purpose handles this by making the experience fun—but not all human computation systems are fun enough to be self-motivating, particularly at the prototyping stage. Mechanical Turk is a good prototyping platform for many forms of human computation, because it offers a ready service for recruiting a crowd on demand. And the first prototypes for VizWiz and Soylent were built on Mechanical Turk.

Yet thinking about programming with human beings inside the system poses special problems. For example with Mechanical Turk, a request for a human to do a small task can take a few minutes and cost a few cents to get a result, which is astounding in one sense (that you can obtain human assistance so quickly and so cheaply), but is abysmally slow and expensive compared to a conventional function call.

"A group of many people has abilities and knowledge that one single end-user does not... The fact is, a crowd comprises a wide range of ideas, opinions, and skills."

Programmers need new tools that can help them experiment with human computation in their systems. For example, our TurKit toolkit [3] integrates Mechanical Turk calls in a traditional imperative/object-oriented programming paradigm, so that programmers can write algorithms that incorporate human computation in a familiar way. TurKit does this using a novel programming model called "crash and rerun," which is suited to long-running distributed processes where local computation (done by software) is cheap, and remote work (done by humans) is costly.

The insight of crash-and-rerun programming is that if our program crashes, it is cheap to rerun the entire program up to the place where it crashed. This is true as long as rerunning does not re-perform all the costly external operations from the previous run. The latter problem is solved by recording information in a database every time a costly operation is executed.

Costly operations are marked by a

Figure 2: In Soylent, after the crowd has suggested words or phrases that can be edited, the end-user can shorten his or her text interactively with a slider. Red text indicates locations where cuts or rewrites have occurred.

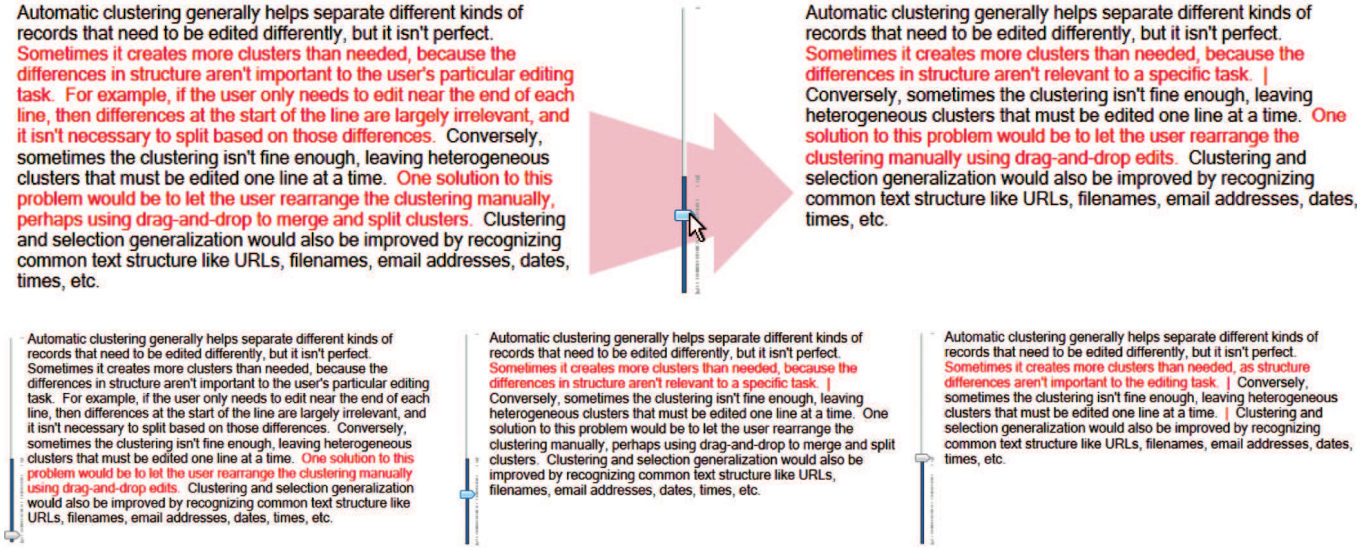
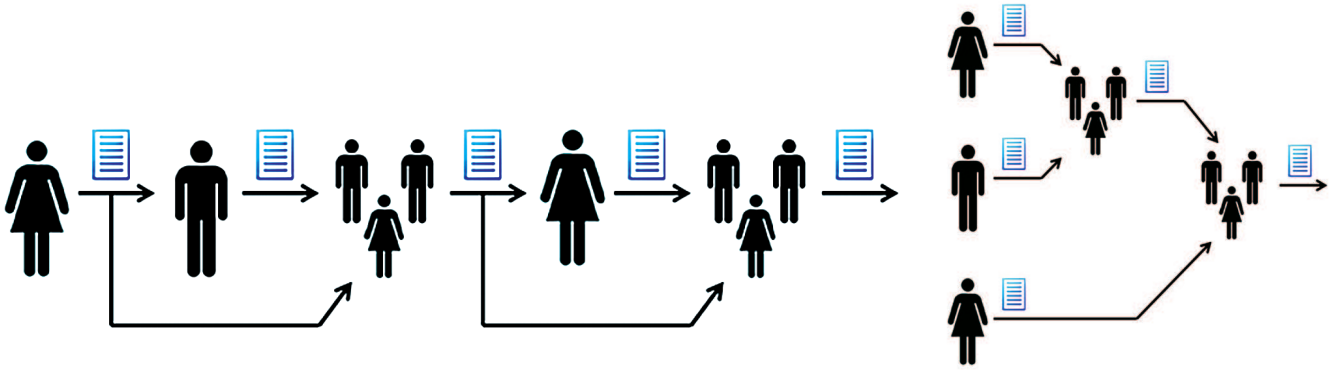


Figure 3: Some human computation processes are iterative (left), involving a succession of interleaved improvement steps (by one person) and voting steps (by several people). Other processes are parallel (right), in which individuals generate original content, and voters simply choose among the alternatives.



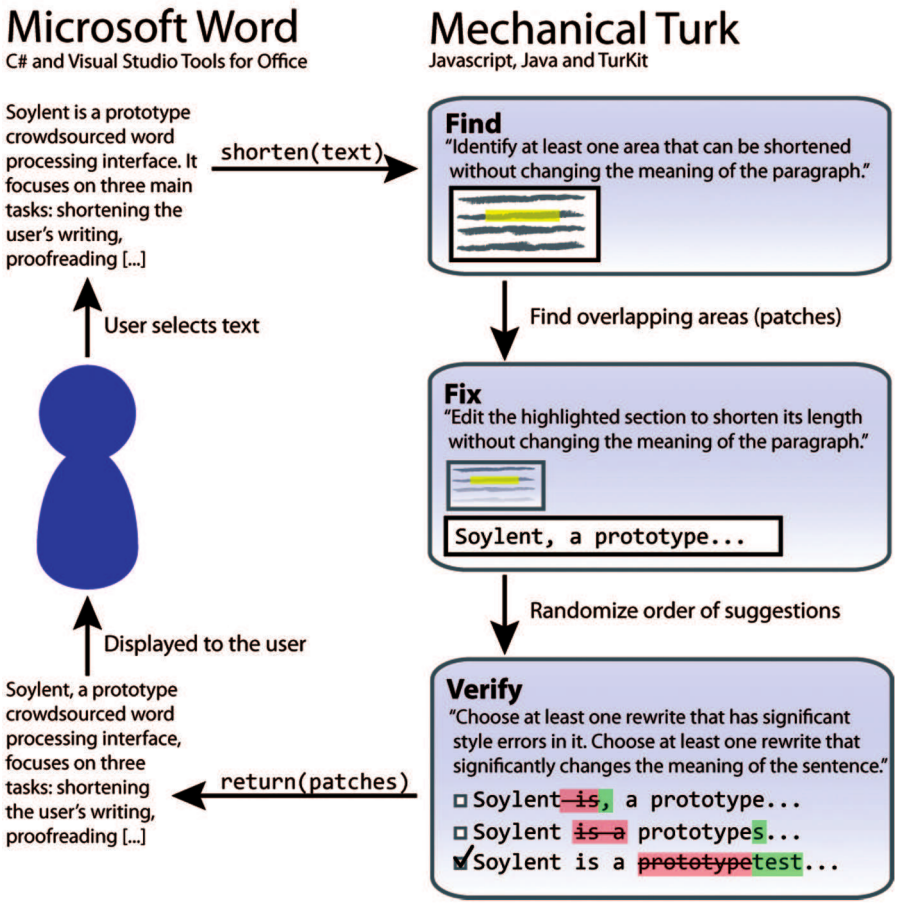
new primitive called "once," meaning they should only be executed once over all reruns of a program. Subsequent runs of the program check the database before performing operations marked with "once" to see if they have already been executed. This model makes it much easier to code algorithms involving human computation. For example, a TurKit program can sort a list of images using human preference judgments by calling the human computation in the sort algorithm's comparison function, and wrapping those calls in "once" to make them persistent.

Another programming challenge is the development of algorithms and design patterns that handle the idiosyncrasies of human beings. Humans are not programmable machines, and they don't always follow instructions, unintentionally or otherwise. Sometimes this should be embraced and supported, to harness the creativity and diversity of the crowd. Other times, it simply produces noisy, erroneous, or useless results.

For example, we have studied alternative algorithms for content creation [4]. Iterative processes are similar to Wikipedia or open-source software development. People build on existing content created by others, with voting or independent review ensuring that the process stays on track. Parallel processes are often seen in design contests, like Threadless.com, where people generate content independently, and then the best is chosen through a vote. See **Figure 3**.

In experiments involving various kinds of work, such as handwriting transcription, image description, and brainstorming, our results show that iterative processes generally produce higher than average quality than parallel processes. However, in the case of brainstorming, workers riff on good ideas that they see to create more good ones, but the very best ideas seem to come from people working alone. And

Figure 4: The find-fix-verify algorithm in Soylent identifies patches in need of editing, suggests fixes to the patches, and votes on those fixes.



with transcription tasks, it turns out that showing workers the guesses of other workers often leads them astray, especially if the guesses are self-consistent but wrong.

Crowd workers exhibit high variance in the amount of effort they invest in a task. Some are lazy turkers, who do as little work as necessary to get paid, while others are eager beavers, who go above and beyond the requirements, either to be helpful or to signal that they aren't lazy turkers, but in counterproductive ways. We need new design patterns for algorithms involving human computation that recognize and control this behavior.

For example, Soylent uses a find-fix-verify pattern to improve the quality of proofreading and document shortening (**Figure 4**). In this pattern, some workers find problems, other workers fix them, and still other workers verify the fixes. But questions remain. What other algorithms and design patterns are useful? How should algorithms involving human computation be evaluated and compared from a theoretical point of view?

SYSTEMS PROBLEMS

Moving from prototyping to actual deployment requires facing questions about how to obtain a reliable and well-performing source of human computation for the system. How can we recruit a crowd to help, and motivate it to continue to help over time, while optimizing for cost, latency, bandwidth, quality, churn, and other parameters?

For paid crowds, these questions intersect with labor economics. Some of our recent work has found that workers in human computation markets like Mechanical Turk behave in unusual ways. For example, instead of seeking work that provides a target wage, they often seek a target earning amount, and simply work until they reach their target, consistent with game-playing behavior [5].

Another difference in these markets is the overwhelming importance of searchability. Workers' ability to find tasks they want to do is strongly affected by the kind of interface the market offers. Mechanical Turk, for example, typically displays a list of thousands of available tasks, divided into hundreds of result pages, with few effective tools for searching or filtering this list. We have found that most workers simply choose a particular sort order and work their way through the list. They most often sort by newest task, or most tasks available, and surprisingly not by price. The speed of completion of a task is strongly affected by its ability to be found, which may not be strongly related to the monetary reward it offers [6].

We can also think about human computation in computer systems terms, such as cost, latency, and parallelism. Services like VizWiz and Sinch need to return answers quickly, and to support that, we have developed an approach (and accompanying implementation) called quikTurkit that provides a layer of abstraction on top of Mechanical Turk to intelligently recruit multiple workers before they're needed.

In a field deployment of VizWiz, users had to wait a little longer than two minutes on average to get their first answer. Wait times decreased sharply when questions and photos were easy for workers to understand. Answers were returned at an average cost per question of only $0.07 for 3.3 answers. Given that other visual-assistance tools for the blind can cost upwards of $1,000 (the equivalent of nearly 15,000 uses of VizWiz), we believe that human computation embedded in an inexpensive software system can be not only more effective but also competitive with, or even cheaper than, existing pure software solutions. When set to maintain a steady pool of workers (at a cost of less than $5 per hour), quikTurkit can obtain answers in less than 30 seconds.

Beyond monetary compensation, many other reasons entice people to participate in human computation, in-

cluding altruism, entertainment, and friendship. How do those motivations influence system performance? And how should the systems be designed to encourage some motivations, and perhaps discourage others?

After demonstrating that VizWiz was feasible using paid strangers on Mechanical Turk, we also ported it to Facebook, so that a blind user's sighted friends can help. We are currently studying how people (at least in this context) choose to trade off the strengths and weaknesses of each service. Mechanical Turk is fast but costs money. Facebook is free, and the user's friends might be more motivated to answer, or even more capable since they know more about the person. On the other hand, the user might be less willing to ask certain personal questions to his or her friends, rather than asking an anonymous Mechanical Turk crowd.

PEOPLE VS SYSTEMS

The gap between what software can do and what people can do is shrinking, but a gap of some sort will exist for a long time. Automatic techniques need to be able to fallback to people when necessary to fill in the gaps, enabling interactions that automatic techniques alone can't yet support and helping us design for the future.

So-called *Wizard of Oz* prototyping (wherein a human is hiding behind the curtain, so to speak) is a venerable technique in human-computer interaction and artificial intelligence that makes an intelligent system (or even a not-so-intelligent one) appear to work even though a software backend isn't ready yet. With platforms like Mechanical Turk and Facebook that make human computation practical, we are now at the point where *Wizard of Oz* is not just for prototyping anymore. We can build useful systems with human power inside, and actually deploy them to real users. These systems will stretch the limits of what software can do, and allow us to find out whether the ideas even work and how people would use them.

In addition, we can collect data from actual system use, like VizWiz queries and photos, that might eventually help to replace some or all of the human power with artificial intelligence. From this perspective, AI would speed up performance and reduce labor costs. But human computation made the system possible in the first place.

Acknowledgements

Ideas and work in this article come from many students and collaborators, including Mark Ackerman, David Crowell, Bjoern Hartmann, David Karger, Marc Grimson, and Katrina Panovich. This work was supported in part by Quanta Computer, NSF, and Xerox.

Biographies

Robert C. Miller is an associate professor of computer science at Massachusetts Institute of Technology. He grew up in rural Louisiana and plays the accordion, although extremely poorly.

Danny "Greg" Little is a PhD student in computer science at MIT. He is probably asleep right now.

Michael Bernstein is a PhD student in computer science at MIT and a features editor for this magazine.

Jeffrey P. Bigham is an assistant professor of computer science at University of Rochester, and he wants to be mayor of your living room. Don't let him in.

Lydia B. Chilton is a PhD student in computer science at University of Washington. When not studying human computation, she is a consultant to the United Federation of Planets and makes the occasional journey on the USS Enterprise with her old pals Kirk, Spock, and McCoy.

Max Goldman is a PhD student in computer science at MIT. When he gives a talk, his performance is so lively and engaging that it distracts the audience from the actual research results, but everybody goes away happy.

John J. Horton is a PhD student in public policy at Harvard University. He is having trouble thinking of a public policy angle for his human computation research and needs to defend his dissertation soon. This is a growing problem that, to date, Turkers have been unable to solve.

Rajeev Nayak is a graduate student in computer science at MIT. He bats .400, shoots .575 from the field, sings *a cappella*, and watches *Glee* religiously.

References

1. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Whyte, B., White, S., Yeh, T. VizWiz: Nearly Real-time Answers to Visual Questions. *UIST 2010*.
2. Bernstein, M., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., Panovich, K. Soylent: A Word Processor with a Crowd Inside. *UIST 2010*.
3. Little, G., Chilton, L., Goldman, M., Miller, R.C. TurKit: Human Computation Algorithms on Mechanical Turk. *UIST 2010*.
4. Little, G., Chilton, L., Goldman, M., Miller, R.C. Exploring Iterative and Parallel Human Computation Processes. *HCOMP 2010*, to appear.
5. Horton, J.J. and Chilton, L. The Labor Economics of Paid Crowdsourcing. *EC 2010*.
6. Chilton, L., Horton, J.J., Miller, R.C., Azenkot, S. Task Search in a Human Computation Market. *HCOMP 2010*, to appear.
7. Nayak, R., *et al*. Sinch: Searching Intelligently on a Mobile Device. *CHI 2011*, in submission.
8. von Ahn, L., and Dabbish, L. Designing Games with a Purpose. *CACM*, 51, 8, August 2008.

ACRONYMS

AMT Amazon Mechanical Turk: a web service owned by Amazon that facilitates crowdsourcing

CAPTCHA complete automated public Turing test to tell computers and humans apart; it's a contrived acronym intentionally redolent of the word "capture," used to describe a test issued on web forms to protect against automated responses

GWAP Game with a Purpose: a computer game that layers a recreational challenge on top of a problem that demands human intelligence for efficient solution, e.g., protein folding

HCIR human–computer information retrieval

HIT human intelligence task: a task that an AMT requester is willing to pay to have accomplished by AMT providers. More generally, a task that may be best completed via crowdsourcing

HuGS human-guided search: A research project investigating a strategy for search and optimization problems that incorporates human intuition and insight

reCAPTCHA a kind of "CAPTCHA" (see above) service that helps to digitize books, newspapers and old radio shows. Used to detect whether a user is a human or a computer (bot)

TF-IDF term frequency-inverse document frequency: A weight formula used in data mining to determine the importance of a particular term to a document in a corpus of text