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 Mechanical Turk ("Turkers") to complete a set of six closed-ended, qualitative content analysis questions using an online survey interface. All subjects in all treatment groups (except one of the two 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 through the survey interface. 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

[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.

choose to leave a blank response.

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 different 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 ended up paying all subjects the largest amount they could have received from their experimental treatment in order to avoid under-paying any deserving subjects.

All control and treatment conditions are described in further detail below. For each of the treatment conditions (listed in bold) we have noted in parentheses whether it is social, financial or hybrid in nature and included the full treatment text. Where appropriate, we have also included 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

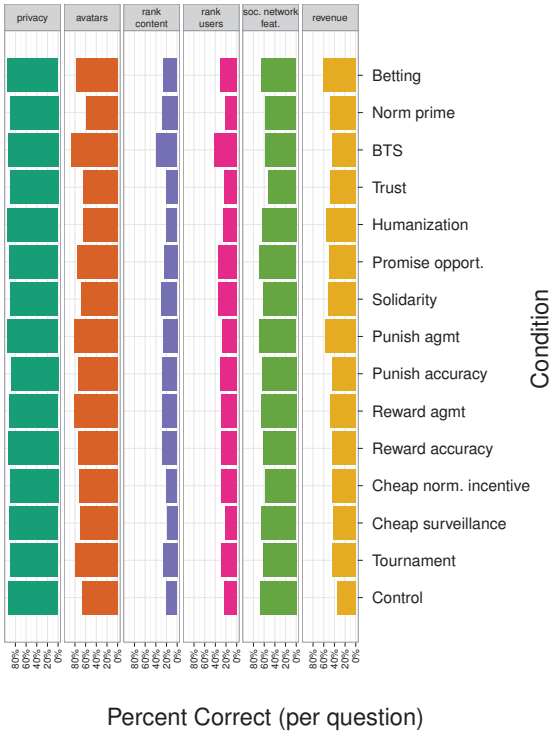

Figure 2. Performance Distribution - All Conditions

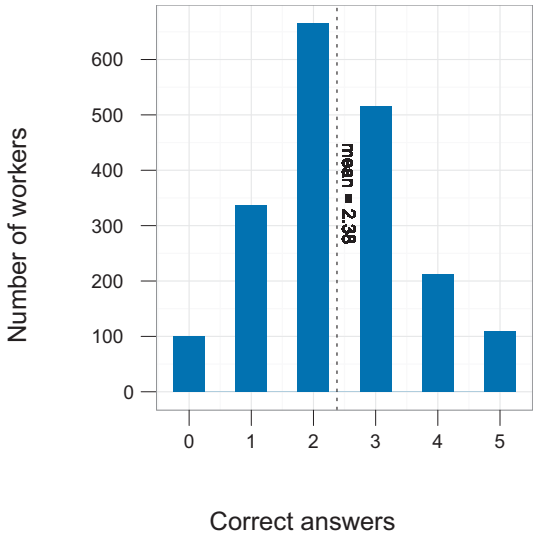

Table 2. Average Treatment Effects (ATE) on Aggregate Performance

	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.

[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.

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

[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 our outcome variable (information seeking task performance). This implies that culturally specific knowledge and experience online may play an important role mediating workers' ability to perform the sort of qualitative information-seeking task we asked them to do here.[17]

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

[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]

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.