Financial Incentives and the "Performance of Crowds"

Winter Mason
Yahoo! Research
111 West 40th Street
New York, NY 10018
+1 (212) 571-8124
winteram@yahoo-inc.com

Duncan J. Watts
Yahoo! Research
111 West 40th Street
New York, NY 10018
+1 (212) 571-8126
djw@yahoo-inc.com

ABSTRACT

The relationship between financial incentives and performance, long of interest to social scientists, has gained new relevance with the advent of web-based "crowd-sourcing" models of production. Here we investigate the effect of compensation on performance in the context of two experiments, conducted on Amazon's Mechanical Turk (AMT). We find that increased financial incentives increase the quantity, but not the quality, of work performed by participants, where the difference appears to be due to an "anchoring" effect: workers who were paid more also perceived the value of their work to be greater, and thus were no more motivated than workers paid less. In contrast with compensation levels, we find the details of the compensation scheme do matter—specifically, a "quota" system results in better work for less pay than an equivalent "piece rate" system. Although counterintuitive, these findings are consistent with previous laboratory studies, and may have real-world analogs as well.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences – *sociology, economics.*

General Terms

Management, Performance, Economics, Experimentation, Human Factors.

Keywords

Incentives, Performance, Crowd-sourcing, Peer Production, Mechanical Turk

1. INTRODUCTION

One of the most exciting and potentially transformative features of the World Wide Web is its ability to connect large numbers of otherwise disparate individuals who wish to contribute to a joint project or community [1-3]. This general movement towards online "peer production," manifests itself in examples as varied as open source software, Wikipedia, and social tagging sites like Flickr and Del.icio.us. One important sub-class of peer production is a phenomenon known as "crowd-sourcing" [4, 5] in which potentially large jobs are broken into many small tasks that are then outsourced directly to individual workers via public solicitation. Workers sometimes work for free, motivated either out of intrinsic enjoyment [3] or some form of social reward [6]; however, successful examples of volunteer crowd sourcing have proven difficult to replicate, in part because arbitrary tasks tend not to be intrinsically enjoyable, and in part because social rewards are often highly context specific. As a result, crowd sourcing increasingly uses financial compensation, often in the form of micro-payments on the order of a few cents per task. This model has the advantage that it is more replicable than models based on intrinsic or social rewards; yet it can still accomplish tasks quickly and cheaply. As a result, paid crowd-sourcing has elicited considerable interest as an alternative mode of production to traditional firms [4]. Nevertheless, the success of any given enterprise still depends to some extent on the ability of would-be "employers" to attract the appropriate workers and motivate them to perform the task well. Although novel in some important respects, therefore, the crowd-sourcing model is faced with a question that has long concerned economists, psychologists, and management theorists; that is, whether and how financial incentives can be used to motivate workplace performance.

Related work. Traditional economic theory has generally espoused the view that rational workers will choose to improve their performance in response to a scheme that rewards such improvements with financial gain [7, 8]. This "rational choice" view is increasingly reflected in management practice—for example, the fraction of executive compensation that is tied to stock price has increased dramatically in recent decades [9, 10]—and has also been supported by a small number of field studies. Most notably, Lazear [8] conducted a study of a large "autoglass" factory in which workers installed windshields on a production line that switched from a time-rate wage (i.e., pay per hour) to a piece-rate (i.e., pay per unit) over the course of a year and a half. Lazear found that individual productivity for workers who started in the time-rate scheme and switched to the piece-rate scheme increased by 20%, leading him to conclude that performance-based pay schemes are a powerful tool for eliciting improved performance.

As many psychologists and management theorists have pointed out, however, results of this kind do not tell the whole story. Numerous experiments have demonstrated that under certain circumstances the provision of financial incentives can undermine "intrinsic motivation" (e.g. enjoyment, desire to help out), possibly leading to poorer outcomes [11, 12]. Alternatively,

KDD-HCOMP '09, June 28, 2009, Paris, France.

workers may ignore rational incentives to work longer when they have accomplished pre-set targets [13]. Even in situations where financial incentives do increase motivation, moreover, recent experiments have demonstrated that they may still undermine actual performance through what is called a "choking effect" [14]. Finally, for more complex tasks, where performance is multifaceted and often hard to measure, performance-based pay schemes can undermine performance in other ways—for example, by encouraging workers to focus only on the aspects of their jobs that are actively measured [15]; to free-ride on the efforts of others [7]; to avoid making colleagues look bad [16], or to avoid taking risks, thereby hampering innovation [17, 18].

The present work. Here we investigate the relationship between financial compensation and performance in two experiments conducted on a particular crowd-sourcing platform, Amazon's Mechanical Turk (see https://www.mturk.com/mturk/welcome). Before describing the experiments in detail, we note that AMT offers some interesting advantages (and some limitations) as an environment within which to conduct experimental behavioral science. In particular, AMT can be used to create a reasonably flexible and lightweight experimental framework that allows experimenters to conduct a wide range of experiments involving potentially large numbers of participants (hundreds or even thousands, but probably not millions) quickly and cheaply. These features are particularly helpful in the current context, where we would like to assign participants to a relatively high number of experimental conditions, as well as check the robustness of findings by varying either the information environment, or even the task itself. Although these manipulations would be possible in a traditional lab setting, by running the experiments on AMT the time and cost is much lower and the results pertain to a real (though atypical) labor market. Nonetheless, the use of an online platform also brings with it certain restrictions with respect to the type of experiments that can be conducted, and raises some novel challenges regarding subject recruitment and retention.

Outline and Main Contributions. The remainder of the paper proceeds as follows. In the next section, we describe the Mechanical Turk platform, and discuss in more detail its advantages and limitations as an experimental platform. In Sections 3 and 4, we then describe two experiments that we conducted on AMT to investigate the relationship between financial incentives and performance. Section 3 describes an experiment in which subjects were asked to sort up to 99 sets of images of moving traffic into their correct temporal order, where subjects were assigned randomly to one of three difficulty levels and one of three rates of pay, yielding nine experimental conditions in all. Section 4 then describes a second experiment, in which subjects solved up to 24 word puzzles, each of which required them to locate target words in a two-dimensional grid of letters, and where they were compensated either on a "per word" or a "per puzzle" basis. Given the distinct nature of the tasks in these experiments, it is not surprising that we found somewhat different results; however, two basic findings seem robust: first, that paying subjects elicited higher output than not paying them (where in the case of experiment one, increasing their pay rate also yielded higher output); and second, that in contrast to the quantity of work done, paying subjects did not affect their accuracy. Although surprising, this latter result may be related to an "anchoring effect" [19-21] in that subjects' perception of the value of their work was strongly correlated with their actual pay rate. In section 5, we discuss the implications of these results and point out some important similarities and differences with the existing literature on financial incentives and performance.

2. Amazon's Mechanical Turk

The original purpose for Amazon's Mechanical Turk (AMT) was to serve as a programmatic interface for tasks that were easier for humans than for machines; however, it can equally be considered a labor market in which "requestors," who can be individuals or corporations, can list tasks (called "human intelligence tasks," or HITs) along with a specified compensation. Individual workers can then elect to complete any number of these tasks for which they are then paid by the corresponding requestor.

When choosing a task to work on, workers are presented with a list of "requests," each of which contains the title of the job being offered, the reward being offered per HIT, and the number of HITs available for that request. Workers can click on a link to view a brief description of the task, or can request a preview of the HIT. After seeing the preview, workers can choose to accept the HIT, at which point the work is officially assigned to them and they can begin completing the task. HITs range widely in size and nature, requiring from seconds to hours to complete, and compensation varies accordingly, but is typically on the order of $0.01-$0.10 per HIT. Currently, several hundred requests may be available on any given day, representing tens of thousands of HITs (i.e. a single request may comprise hundreds or even thousands of individual HITs); thus while AMT is only one particular instantiation of the crowd sourcing model, its size and diversity make it an attractive object of study.

As we have already noted, the Mechanical Turk framework can also be thought of as a convenient pool of subjects willing to participate in laboratory-style behavioral experiments for a relatively low fee (where the nature of the experiments are appropriately disclosed). For the specific research question at hand—the relationship between financial incentives and performance—a major advantage of adopting an experimental approach is that it allows us to eliminate many of the confounding effects that arise in real-life employment contexts, such as free-riding, risk-avoidance, or group interaction effects. The degree of control over the task, setting, and incentive structure also allows us to restrict attention to a single aspect of the overall relation between financial incentives and performance—namely whether simply increasing the rate of compensation for a given task leads to better performance.

Generalizing results from the crowd-sourcing context to the offline context requires caution, however, as there are at least two potentially important differences between the two contexts; thus one might suspect that studies conducted on AMT may lack external validity. The first difference is that the highly self-selected AMT population may not be representative of the population at large, both because it is an exclusively online environment, and also because of the unconventional nature of "labor" being provided. Fortunately, one advantage of our particular study design is that individuals are randomly assigned to different payment conditions; thus whatever differences we observe across conditions is attributable to the conditions themselves, net of whatever selection biases are responsible for people participating in the first place. Also, as we show below, our subject pool is surprisingly diverse, consistent with previous studies.

The second issue is that payments in crowd-sourcing are much smaller (cents per task) than would typically be the case in lab experiments, and trivial compared with traditional labor markets. One might therefore suspect that participants in our experiments will not respond in a sensible way to incentives because they are motivated primarily by non-financial incentives, or are simply not taking the work seriously. As we discuss later, the issue of motivation may indeed pose some problems for external validity, and certainly invites further study. We note, however, that recent research has indicated that the quality of work performed on AMT is as good, and maybe even better than, work performed by experts paid under traditional contracting arrangements [22], indicating that it is being taken seriously. Moreover, and more importantly for our purposes, we show that in at least some instances subjects do in fact respond sensibly to wage differences, suggesting that there is some external validity to the effects observed in crowd-sourcing contexts.

3. STUDY 1: IMAGE ORDERING

3.1 Methods

To understand the impact of compensation on performance, we wish to differentiate between the quantity of work performed (output) and the quality of the work (accuracy); thus we require a task for which output can vary widely and accuracy can be measured objectively. To meet these criteria, we created a task in which participants sorted a set of images taken from a traffic camera at 2-second intervals into chronological order.

3.1.1 Design

To participate in the study, participants were required to have an account on Amazon's Mechanical Turk. These accounts are associated with a unique ID (which is, in turn, associated with a bank account), so it is possible to ensure each account (and therefore most likely, each person) only participates once. The study, which was listed among other tasks posted by other requesters, had the title, "Reorder traffic images," and was described as "Sort images from traffic cameras in chronological order", with a base rate pay of $0.10. If examined further, participants could see a preview of the HIT (Figure 1), which included a description of the task and an example of the images to be sorted along with the correct order. Some participants were informed that they would be paid an additional bonus for each set of images sorted, and for others there was no such indication. If participants accepted the task, they were asked (but not required) to provide some demographic information, and then were given a chance to practice sorting images. All participants were paid an initial fee of $0.10 to complete the introductory survey and training set, and all received the same three practice sets of three images, displayed in the same (incorrect) order. To sort an image set, participants clicked on the photo they wished to reorder and dragged it into the correct position. When they felt the images were correctly sorted in chronological order from left to right, they pressed a button to submit their sorting and proceed to the next image set. During the practice trials, after submitting each image set they received feedback on whether the images were correctly sorted, and if not, what the correct order was. Participants were informed that the feedback would only be available during the practice trials.

After completing the practice trials, participants were given information about how much they would be paid, and were randomly assigned to a difficulty level: easy (2 images per set), medium (3 images) and hard (4 images), and to one of four pay levels: no pay, low pay ($0.01 per image set), medium pay ($0.05 per image set); and high pay ($0.10 per image set). They were also informed that by continuing, they were consenting to participate in an experiment; however, there was no indication that the difficulty or payments, either the manner or the amount, would be different for anyone else engaging in the task. Participants could sort any number from 0 to 99 sets of images, where the number of image sets they chose to sort was our measure of quantity, and their accuracy in sorting the images was our measure of quality. At any point, they could choose to finish and accept the bonus (if any) for the tasks so far completed. Once they chose to finish sorting image sets, or if they sorted all 99 image sets, they were asked to complete a brief questionnaire about their performance. After this, the participants were given feedback about the number of tasks completed and could submit their work to receive payment.

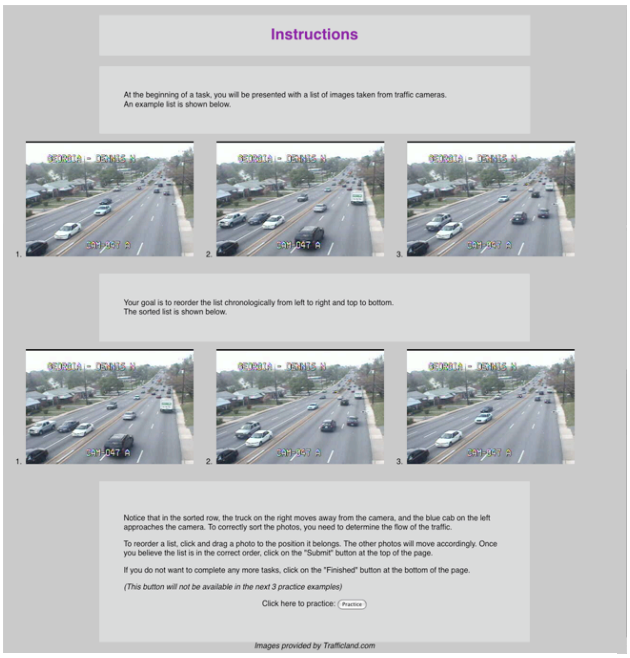

Figure 1. Screenshot of the description of Study 1, which participants saw when deciding to participate in the experiment.

3.1.2 Participants

Over all conditions, the experiment involved 611 participants, who sorted a total of 36,425 image sets. Participants were asked to report their current location, and 594 participants identified 43 different countries. The majority (82.7%) was from the United States, and the next four highest responding countries were India (6.4%), Canada (1.5%), Vietnam (1.2%), and United Kingdom (1%) Asked to report their gender, 563 participants responded, and of these 58.8% were female and 41.2% were male. Of the 568 reporting age, the average response was 33.3 years and the median age was 31 years. Participants were given a choice of five income levels to report, and of the 598 who offered the information, 18.6% reported an income less than $7000, 22.6% reported an income between $7 - $30k, 34.5% reported an income between $30 - $70k, 21.1% reported an income between $70 -

$160k, and 3.2% reported an income greater than $160k. The subject pool was therefore reasonably diverse, consistent with previous user surveys of the AMT population—for example, http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html describes a similar income and gender distribution, and also recorded 58% female respondents.

3.2 Results

Figure 2 reveals two main findings: first, that across all difficulty levels participants chose to complete more tasks on average when the pay was higher ($F(3,607) = 15.73$, $p < 0.001$); and second, that across all payment levels, the number of completed tasks decreased with increasing difficulty. We also observe, however, that there is no interaction between difficulty and compensation, thus hereafter we focus on the effect of pay on quantity averaged over difficulty levels. In addition to the average effect of pay, we also found that more of the participants paid $0.10 sorted the maximum possible than those paid $0.01 or nothing at all, and proportionately more of the participants paid $0.01 sorted fewer than 10 sets. These results, in other words, are completely consistent with standard economic theory, which predicts that the more a person is paid to do X, the more of X they will do [7, 8]. Nevertheless, the finding is reassuring since, as noted above, one might have expected that variability in intrinsic motivation (e.g. enjoyment of the task) would have overwhelmed the effect of changes in extrinsic motivation (payment), which can vary by at most $0.10 per task. The strong and significant dependence of output on compensation therefore suggests that the range of wage rates studied ought to be sufficient to observe variability in the quality of performance as well.

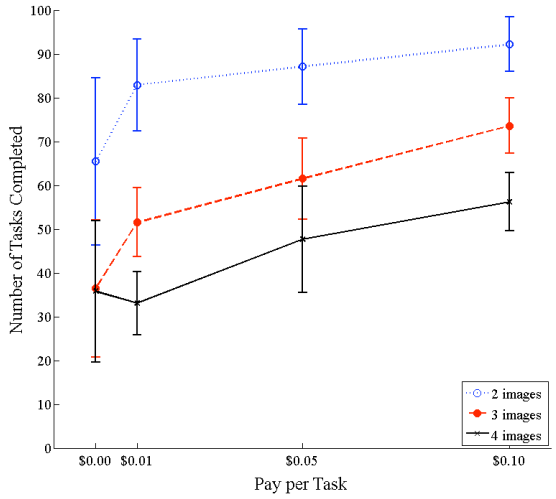

Figure 2. Number of image sets sorted in Study 1 increases with wage rate and decreases with difficult of task; error bars are standard error.

As Figure 3 indicates, however, increasing compensation did not improve accuracy, which we measured in two ways: first, using the proportion of image sets that were sorted into the correct order; and second, using Spearman's rank correlation (ρ), which is the normalized sum of squared differences between the correct order and the sorted order. For each accuracy measure, we confirmed quantitatively what is visually apparent in Figure 3 in two ways: first, using a simple one-way analysis of variance; and

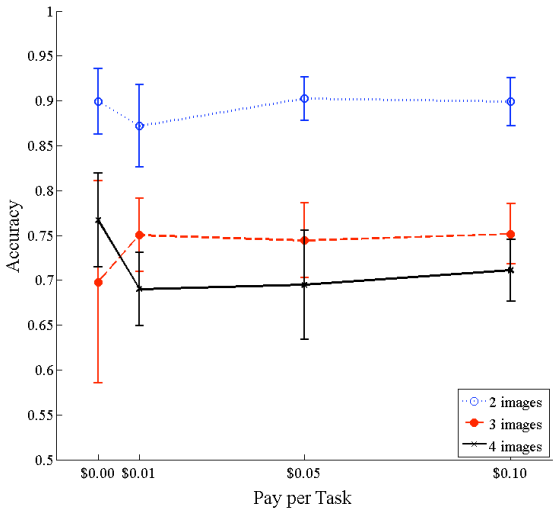

Figure 3. Accuracy, defined as the proportion of image sets correctly sorted, is not reliably different for different wages; error bars are standard error.

second, fitting the data to a hierarchical linear model (where again we averaged our results over the three difficulty levels). Although the number of tasks each participant completed can only be analyzed at the participant level, the measures of accuracy can also be analyzed at the task level; thus, the hierarchical linear model [23], also known as a multi-level model, is a useful statistical model because it accounts for the variability in the inherent difficulty of sorting each image set, and the variable number of tasks each participant completed. In this analysis, the compensation offered is treated as a categorical variable and modeled as a random effect simultaneously with user-level effects and task-level effects. Specifically, when accuracy is defined as the probability $\Pr(y_i = 1)$ that the image set i was sorted correctly, the model is

$$\Pr(y_i = 1) = \text{logit}^{-1}(\alpha_{t[i]} + \beta_{t[i]} \cdot pay_i + \eta_{u[i]}),$$

and when accuracy is defined as Spearman's rank correlation (ρ) between the actual and correct ordering, the model is

$$\hat{\rho} = \alpha_{t[i]} + \beta_{t[i]} \cdot pay_i + \eta_{u[i]},$$

where in both cases $\alpha_{t[i]}$ is the intercept for each task, $\beta_{t[i]}$ is the slope for the wage received, and $\eta_{u[i]} \sim N(0,\sigma)$ is the intercept for each user. The parameters α and η therefore capture variance among different tasks and respondents respectively, and β captures the effects of the wage rate.

Table 1. Average parameter estimates for the effect of pay in the hierarchical linear model across users.

	Model estimate $(\bar{\beta})$	95% Confidence Interval
$0.00	0.059	(-0.055, 0.173)
$0.01	-0.124	(-0.220, -0.029)
$0.05	-0.057	(-0.154, 0.041)
$0.10	0.086	(-0.0044, 0.1775)

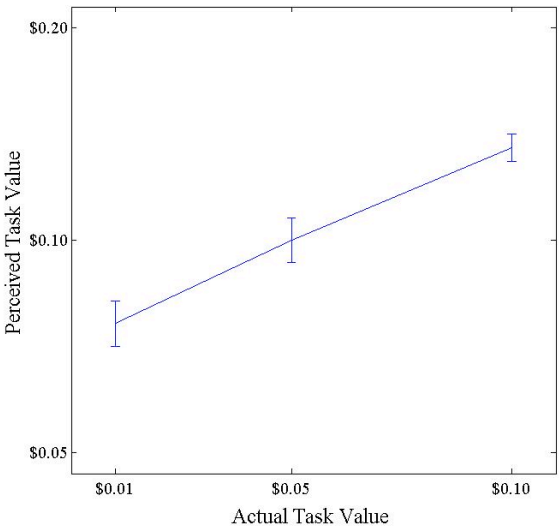

Figure 4. Post-hoc survey shows perceived value of the task increases with the actual pay, but is always slightly greater than the actual pay received.

Regardless of the accuracy measure or analytical method used, we found that the wage rate had no significant effect on the participants' accuracy in sorting the image sets. First, as indicated in Table 1, the parameter estimates in the hierarchical model for the four levels of pay were not reliably different from each other; and second, one-way ANOVAs of wage rate on proportion correct and rank correlation were not statistically reliable (proportion correct: $F(3,607) = 0.66$, *ns*; rank correlation: $F(3,607) = 0.82$, *ns*).

3.3 Discussion

One possible explanation for the absence of an effect of wages on accuracy is that subjects simply assumed they would be paid regardless of performance. This explanation is somewhat unlikely, as AMT's policy is that requestors are only obligated to pay for accurate or useful work, and workers are informed of the policy. Nevertheless, to check the possibility we ran an additional experiment with a single payment level ($0.01) that provided different information to participants regarding the importance of accuracy. In this additional experiment, some participants were given the same instructions as before while others were told that one out of every four image sets was a test image set used to gauge their accuracy. Within this latter condition, we also created four variants: (i) participants only informed that accuracy would be measured; (ii) participants also shown feedback on their accuracy after every fourth image set; (iii) participants also told explicitly that their pay would be contingent on their performance; and (iv) participants shown feedback and also told that pay was contingent. We found that quantity and quality results were indistinguishable in all these conditions, suggesting that participants in all conditions were in fact treating their pay as performance dependent.

Although the differential effect of pay on quantity and quality is at first puzzling, we note that previous studies have also found positive effects of financial incentives on quantity of work performed but no effect on quality [24]. We hypothesize, moreover, that the difference derives from an "anchoring" effect, similar to effects that have been observed in other domains of judgment and decision-making [19-21]. As Figure 4 shows, when surveyed after the completion of their tasks, workers in all conditions generally felt that the appropriate compensation for the work they had just performed was greater than what they had received, but the values they expressed depended significantly ($\chi^2 = 243.61$, $p < 0.0001$) on their actual compensation: on average, workers paid $0.01 per task felt they should have received $0.05; workers who were paid $0.05 felt they should have received $0.08; and workers who were paid and $0.10 felt they should have received $0.13. On the one hand, therefore, paying people more to perform a task makes that task more attractive relative to their available outside options, such as other HITs on AMT; thus subjects in the higher pay conditions stayed longer and completed more tasks than those in low pay condition. On the other hand, because of the anchoring effect, all workers felt like they were being paid less than they deserved; thus were no more motivated to perform better no matter how much they were actually paid.

4. STUDY 2: WORD PUZZLES

4.1 Methods

In spite of this explanation, one might suspect that the absence of an effect on accuracy may be an artifact of the task itself—because, for example, it allowed only a small number of potential solutions (in the "easy" condition, for example, only two solutions were possible); or because subjects could not easily improve the quality of their answers with greater effort. To address this possibility, we performed another experiment, using a similar experimental design, but changing the task to finding words hidden in a random array of letters (see Figure 5).

4.1.1 Design

For each puzzle, we provided a list of words that might be found in the puzzle, although only a subset of the list was actually hidden in the word puzzle. As before, this task allowed us to measure quantity (number of puzzles completed) and quality (fraction of words found per puzzle) independently; but because participants did not know how many words from the list could

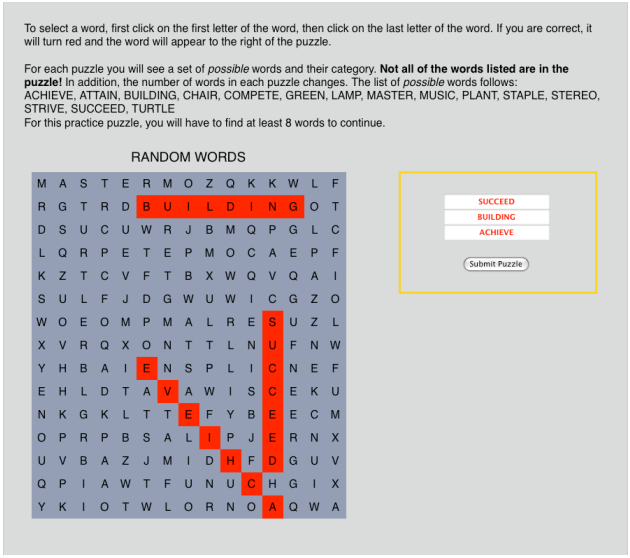

Figure 5. Screenshot of Study 2. Participants found words hidden in a grid of letters.

actually be found in the puzzle, they could not be certain that they had found all the words. Also as before, we randomly allocated participants to different experimental conditions, where in addition to varying compensation levels we also created two distinct compensation schemes [cf. 25]—a "quota" scheme, in which individuals were told they would be paid for every puzzle successfully completed, and a "piece rate" scheme in which they were paid for every word found. Within each compensation scheme, we once again created three pay levels—low, medium, and high—as well as a no pay condition. In total, therefore, Experiment 2 comprised 7 distinct experimental conditions.

The recruitment of participants was the same as in Study 1, with the exception that the title was "Solve Word Jumble puzzles," and the description now read, "Try to find words hidden in a jumble of letters." If participants previewed the HIT, they saw a description of the task that included a small example puzzle and a brief description of the task. If participants accepted the task, they were asked to complete a demographics questionnaire, and then were given more detailed instructions about how to do the task. For each puzzle, there was a list of 15 words, some of which could be found in the array of letters. The puzzles had between 7-14 hidden words, with a median of 11 hidden words per puzzle. To select a word, participants would click on the start and end of a word, after which the word would appear in a panel on the right and the word would remain highlighted (see Figure 5). Once they felt they had found all of the hidden words, they could click a button to continue to the next puzzle.

Participants would encounter one of two practice puzzles, each with 12 hidden words. If participants tried to continue before successfully finding at least 8 of the words, they were asked to continue finding words. Once they had found at least 8 words in the practice puzzle, they received feedback informing how many words they could have found and how many they had found. At this point they were also informed how much they would be paid (if at all), whether the payment would be by puzzle or by word, and that if they continued they were giving their consent to participate in an experiment. Again, there was no indication that the manner or amount of pay would be any different for anyone else. Participants were informed that they could complete as many puzzles as they liked, up to a maximum of 24 puzzles. If the participant chose to continue, they could then take as much time as they wanted on each puzzle, and once they felt they had found all of the words, they could move on to the next puzzle. At any point they could choose to finish and collect their payment. If they chose to finish, or if they completed all 24 puzzles, they were asked to complete a brief post-task questionnaire, and then were given feedback on their performance and could submit their work for payment.

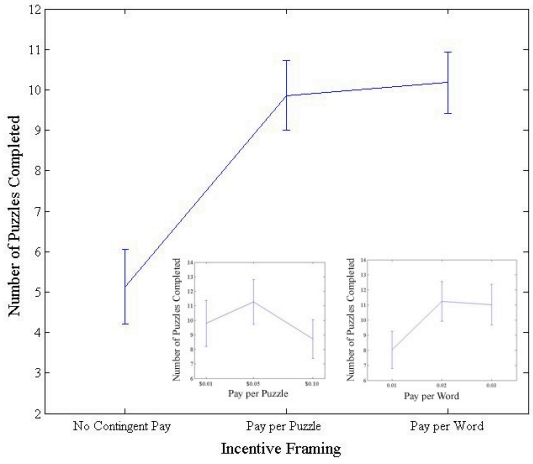

Figure 6. Participants who were paid for each puzzle completed or word found completed significantly more puzzles than those who did not receive contingent pay. Insets show number of puzzles completed did not differ by pay, within per-puzzle (left) or per-word (right) schemes.

4.1.2 Participants

Over all conditions, 320 participants solved a total of 2736 puzzles, finding 23,440 words. Participants were asked to report their current location, and 309 participants identified 19 different countries. The majority (83.9%) was from the United States, and the next four highest responding countries were India (4.8%), Philippines (2.5%), Canada (1.9%), and United Kingdom (1.3%). Of the 303 ages reported, the average was 34.6 years and the median age was 32 years. Participants were again given a choice of income levels to report, and of the 303 reporting, 6.8% reported an income less than $7000, 30.7% reported an income between $7 - $30k, 40.8% reported an income between $30 - $70k, 19.7% reported an income between $70 - $160k, and 1.9% reported an income greater than $160k. These self-reported descriptive statistics, in other words, were generally consistent with those from Study 1, which encourages us to believe that they are reliable. A striking difference with Study 1, however, was that of the 290 participants who reported their gender 74.1% were female, as opposed to 58.8%. Although we can only speculate about the reason for this disparity, one possible explanation is the task itself—that is, women may enjoy completing word puzzles more than men, whereas image sorting is more equally appealing (or unappealing) to both genders. If true, the importance of intrinsic enjoyment in task selection raises the concern from earlier that it may also undermine the impact of financial incentives on task completion—a concern that as we show next, appears to be valid.

4.2 Results

As with the first experiment, we found that effort-contingent pay motivated participants to do more work: participants who were paid either on a quota or a piece-rate basis completed more puzzles (Figure 6) and found more words than participants who were not paid, $F(2,303) = 8.72$, $p < 0.001$. Looking within the two compensation schemes, however, we found no significant impact of compensation on quantity of work (see Figure 6, insets)—a notable difference from the previous experiment (per puzzle: $F(2,108) = 0.71$, *ns*; per word $F(2,124) = 1.82$, *ns*). Why the level of compensation did not have an effect is not clear, but it is likely that intrinsic motivation may have played a larger role in this task than the previous one, as indeed is indicated by the strong bias towards female participation. Even more striking, one participant in the unpaid condition spent five hours completing all 24 puzzles, and found all but 2 words, for a total compensation of $0.10. There was also a very strong relationship between the participants' post-hoc rating of their enjoyment of the task and the number of puzzles they completed, $F(5,299) = 11.06$, $p<0.001$; those who enjoyed it the most completed 19.3 puzzles on average, compared to the 6.2 puzzles completed by those who only enjoyed

Table 2. Average parameter estimates for the effect of pay in the hierarchical linear model across users when A) paid per puzzle; and B) paid per word.

2A	Model estimate $(\bar{\beta})$	95% Confidence Interval
$0.01	-0.046	(-0.066, -0.027)
$0.05	0.036	(0.016, 0.056)
$0.10	0.01	(-0.01, 0.03)

2B	Model estimate $(\bar{\beta})$	95% Confidence Interval
$0.01	-0.022	(-0.0464, 0.0027)
$0.02	0.036	(0.015, 0.056)
$0.03	-0.014	(-0.0357, 0.0075)

it a little. Presumably, therefore, at least some participants found the task intrinsically enjoyable, thus diminishing the impact of extrinsic motivation. Nevertheless, it remained the case that paying participants to work generated more work than not paying them, suggesting that although extrinsic motivation play a less important than in task 1, it remains relevant to output.

Given this weaker relationship between pay and quantity of work, it is not surprising that we again found that the level of compensation had no significant effect on the quality of performance—measured here as the fraction of total possible words found per puzzle—within either scheme. This was true whether we fit the participants' average accuracies across puzzles using an ANOVA, or fit the data to a hierarchical linear model (Eq. 1) that accounted for the variability in difficulty across puzzles and number of puzzles completed (see Table 2A & 2B).What was surprising, however, was that the compensation scheme itself had a large effect on accuracy. Figure 7 (solid line) shows the per-word equivalent pay for the three payment schemes: no pay ($0.008 per word); per puzzle (average of $0.011 per word); and per-word ($0.025 per word). Participants in the "pay-per-word" condition, in other words, earned roughly four times as much per word, on average, as participants in the "pay-per-puzzle" condition. Intuitively, therefore, one would expect that participants being paid per-word would find more words per puzzle than those being paid per puzzle, who would in turn find more words than those not being paid at all. As Figure 7 (dashed line) shows, however, the actual ranking was precisely the opposite: participants who did not receive any contingent pay found the most words per puzzle on average (85.6%), while those paid per puzzle found the next highest (84.7%), and those paid per word found the least (81.4%).

4.3 Discussion

Although counterintuitive, the higher work-to-pay ratio for the per-puzzle condition is consistent with previous work [25], which has found that quota systems (analogous to our per-puzzle condition) elicit more effort than piece-rates (i.e. per-word payment). Following this work, we note that the presence of a quota may elicit work in two ways: first, by encouraging greater marginal effort for hard-to-find words that may complete a puzzle; and second, through implicit goal setting (i.e. where completing the puzzle becomes, in effect, a more salient goal). In other words, those in the pay-per-word condition may have

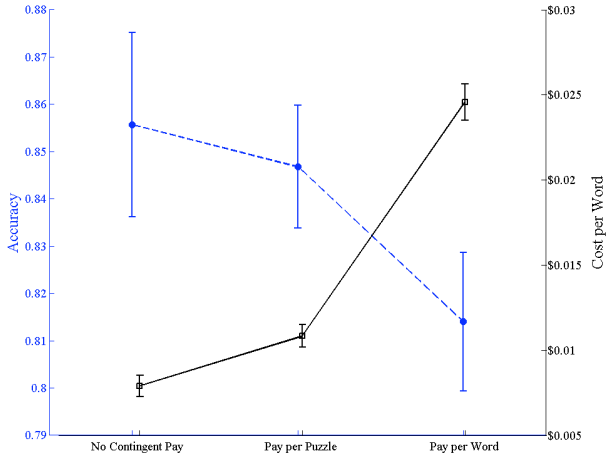

Figure 7. Participants were paid more in total (solid line) but found fewer words per puzzle (dashed line) in the pay per-word condition than in the pay per-puzzle condition.

chosen to advance to the next puzzle when the perceived marginal difficulty of finding the next word in the current puzzle became too great. We note, however, that participants tendency to skip words was not affected by increasing pay, as one would expect if they were making a rational tradeoff between time and compensation. The main impact of the different compensation schemes therefore seems to be psychological, not economic.

Emphasizing this last point, we found a similar anchoring effect with respect to perceived value of the work that we identified in the first experiment (Figs. 9A & 9B). Unlike in the first experiment, however, we also asked participants who received no contingent pay to estimate the value of each puzzle or each word. Surprisingly, these uncompensated participants perceived the task to be worth more than what those paid the lowest non-zero amount perceived it to be worth, but less than those at the highest compensation level. Although at face value this result seems to contradict previous results showing that unpaid volunteers exert more effort than those paid a low wage [12], the difference can be explained by considering the expectations of the participants. When there is no expectation of financial reward, effort is motivated by other kinds of (e.g., social) rewards; but when monetary compensation is expected, as in the AMT framework, the anticipated financial value of the effort will be the driving mechanism. To summarize, therefore, we find although paid workers generally did more work than unpaid workers, how they were paid had a larger impact on their output and accuracy than how much they were paid. Moreover, paying workers a low rate led to them to perceive their work as less valuable than not paying them at all.

5. GENERAL DISCUSSION

In this paper, we have investigated the relationship between financial incentives and performance in the novel setting of online peer production systems. Our main findings are twofold: first, we found that increased payments increased the quantity of work performed, but not its quality; and second, that the particular design of the compensation scheme (a quota scheme versus a piece rate, for example) can have a significant effect on quality,

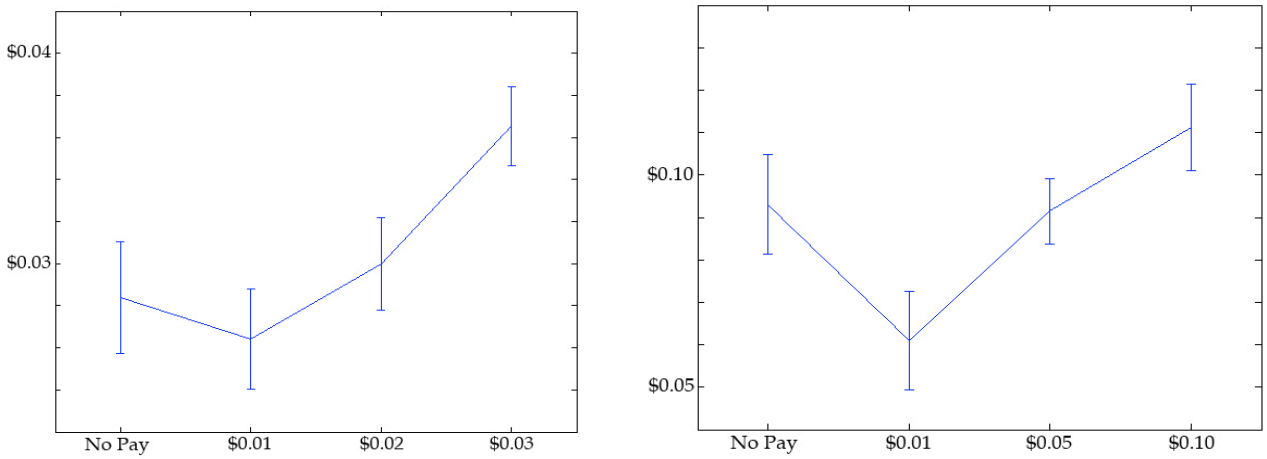

Figure 9. Perceived value of completing a puzzle (left) or a word (right) as reported in a post-task survey relative to the actual pay received. Participants anchored on the value earned, and on average valued the task slightly greater than the amount received. When not given task-contingent pay, participants had no anchor and perceived the value of a task to be higher than those receiving the lowest contingent pay.

even to the point where better work can be accomplished for less pay. A key psychological element to understanding these results appears to be that individuals anchor their perception of what they should be paid directly off of what they are being paid. Moreover, when they are not being paid anything, they have nothing to anchor on, leading to a surprising non-monotonic relation between actual pay and perceived value. Although a number of our results do seem sensitive to the specific nature of the task (e.g. the dependency of quantity on pay rate), this anchoring effect seems to be robust, suggesting that it may be quite general.

These results may have implications for would-be employers wishing to take advantage of crowd-sourcing platforms like AMT. First, when it is possible to use non-financial rewards, such as harnessing intrinsic motivation [2], the quality of the work will be as good or better than using financial rewards, and therefore work can be accomplished as effectively for little to no cost. Second, when it is not possible to incentivize work through intrinsic motivation (i.e., enjoyable tasks) or through social rewards, it may be in the employer's best interest to offer as little as possible—assuming, of course, a large enough crowd exists to make up for the diminished quantity of individual output the low pay would garner. Offering greater reward, in other words, may get the work done faster, but not better.

To what extent these findings generalize beyond the web-specific context of crowd sourcing is more speculative, in part because of the subject pool and pay rate issues raised earlier, and in part because our experimental design skirts some important aspects of financial incentives systems that have been studied elsewhere. For instance, it may be the case that financial incentives exert much of their impact through sorting effects [8]—that is, by offering higher wages or pay that is tied to performance, employers attract and retain better workers. Because we allocate individuals randomly to different pay conditions, we cannot observe any such effects. In addition, discussions about incentives often focus on the contrast between fixed wages and performance-based pay; yet because our experiment considers only the differences between various performance-based pay conditions, our results are silent on this important issue. Obviously, we omitted these effects deliberately in order to focus on the simpler and more specific issue of variable pay rates; however, the narrow focus also increases the difficulty of finding analogues in realistic, offline contexts—that is, cases where indistinguishable workers are paid different rates to do the same work.

Acknowledging the speculative nature of the exercise, however, one provocative analog is the observation that chief executives in Europe are paid considerably less than their American counterparts [26]. It is hard to argue that this disparity exists because Europeans are less talented, work less hard, or that their performance is systematically worse. Rather, it appears to derive instead from historical differences in cultural norms, which have the effect of setting an "anchor," relative to which individuals are judging the value of their work. Analogous to what we see in our experiments, it appears that the particular value of the anchor itself has little effect on individual performance—a point that might also be made about the well-documented pay gap between male and female workers in the US [27] which again appears uncorrelated with actual performance. Finally, a recent study of US Federal circuit judges found that in spite of considerable variation in salary across different states, performance bore no systematic relation to compensation [28], suggesting once again that absolute pay rates are less important to performance than perceptions of relative value.

In addition to these real world analogs, we have already noted that our results bear considerable resemblance to previous experimental findings that have been obtained in physical laboratories where the sums of money at stake were considerably larger than in our case (albeit still small compared with pay in traditional labor markets). In particular, it has been shown that increased financial compensation tends to yield more, but not better work [24], and that quota systems can outperform piece-rates [25]—similar to our findings regarding per-word and per-puzzle payments in Study 2. Although many more experiments of the kind we have described here would be needed to make firm generalizations, therefore, these results do at least suggest that the principles relating compensation and performance may be sufficiently general to span very different contexts and

compensation levels, and thus can be investigated usefully even when payments are very small.

Finally, we note that the fast and economical nature of AMT may make it of interest to behavioral scientists more generally, as an environment for conducting behavioral studies and experiments [22]. Naturally, only web-based studies can make use of this approach, ruling out those that require in-person interactions, physiological measurements, and so on; however, many studies of interest could be run online, including surveys [29], reaction time studies [30], group interaction studies [31] and categorization experiments [32]. Clearly web-based approaches also present novel challenges associated with recruitment bias, participant dropout, etc. [33]; however, we are optimistic that these issues can be addressed, and in some respects the web permits broader and more representative participation than the traditional pool of university students. Crowd-sourcing, in other words, is not only an interesting phenomenon in management and business [4], but could become a useful tool for studying questions of interest to behavioral and social scientists as well.

6. ACKNOWLEDGMENTS

Our thanks to Sharad Goel for helpful feedback, and to Prasenjit Sarkar and Tejaswi Kasturi for assistance in setting up the experiments.

7. REFERENCES

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

[2] Malone, T. W., Laubacher, R. and Dellarocas, C. *Harnessing Crowds: Mapping the Genome of Collective Intelligence*. MIT, City, 2009.

[3] von Ahn, L. Games with a purpose. *Computer*, 39, 6 2006), 92-94.

[4] Howe, J. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business, New York, 2008.

[5] Kleeman, F., Voss, G. G. and Rieder, K. Un(der)paid Innovators: The Commercial Utilization of Consumer Work through Crowdsourcing. *Science, Technology & Innovation Studies*, 4, 1 2008), 5-26.

[6] Nov, O., Naaman, M. and Ye, C. What drives content tagging: the case of photos on Flickr. In *Proceedings of the Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (Florence, Italy, 2008). ACM, [insert City of Publication],[insert 2008 of Publication].

[7] Prendergast, C. The provision of incentives in firms. *J. Econ. Lit.*, 37, 1 1999), 7-63.

[8] Lazear, E. P. Performance pay and productivity. *American Economic Review*, 90, 5 (Dec 2000), 1346-1361.

[9] Murphy, K. J. *Executive Compensation*. SSRN, 1998.

[10] Hall, B. and Liebman, J. B. *Are CEOs Really Paid Like Bureaucrats?* SSRN, 1997.

[11] Gneezy, U. and Rustichini, A. Pay enough or don't pay at all. *Q. J. Econ.*, 115, 3 2000), 791-810.

[12] Heyman, J. and Ariely, D. Effort for Payment: A Tale of Two Markets. *Psychological Science*, 15, 11 2004), 787-793.

[13] Camerer, C., Babcock, L., Loewenstein, G. and Thaler, R. Labor supply of New York City cabdrivers: One day at a time. *Q. J. Econ.*, 112, 2 1997), 407-441.

[14] Ariely, D., Gneezy, U., Loewenstein, G. and Mazar, N. Large Stakes and Big Mistakes. *Review of Economic Studies*Forthcoming, 2008).

[15] Holmstrom, B. and Milgrom, P. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, 71991), 24-52.

[16] Bandiera, O., Barankay, I. and Rasul, I. Social Preferences and the Response to Incentives: Evidence from Personnel Data*. *Q. J. Econ.*, 120, 3 2005), 917-962.

[17] Kohn, A. Why Incentive Plans Cannot Work. *Harvard Business Review*September-October, 1993), 54-63.

[18] Herzberg, F. One More Time: How do You Motivate Employees? *Harvard Business Review*September-October, 1987), 5-16.

[19] Tversky, A. and Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185, 4157 1974), 1124-1131.

[20] Chapman, G. B. and Johnson, E. J. The limits of anchoring. *Journal of Behavioral Decision Making*, 7, 4 1994), 223-242.

[21] Ariely, D., Loewenstein, G. and Prelec, D. Coherent Arbitrariness: Stable Demand Curves Without Stable Preferences. *Q. J. Econ.*, 118, 1 2003), 73-105.

[22] Kittur, A., Chi, E. H. and Suh, B. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (Florence, Italy, 2008). ACM, [insert City of Publication],[insert 2008 of Publication].

[23] Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge University Press, Cambridge, UK, 2006.

[24] Jenkins, G. D., Mitra, A., Gupta, N. and Shaw, J. D. Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology*, 83, 5 1998), 777-787.

[25] Bonner, S. E., Hastie, R., Sprinkle, G. B. and Young, S. M. A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research*, 122000), 19-64.

[26] Conyon, M. J. and Murphy, K. J. The Prince and the Pauper? CEO Pay in the United States and United Kingdom. *The Economic Journal*, 110, 467 2000), F640-F671.

[27] Blau, F. D. and Kahn, L. M. Gender Differences in Pay. *The Journal of Economic Perspectives*, 14, 4 2000), 75-99.

[28] Choi, S. J., Gulati, G. M. and Posner, E. A. Are Judges Overpaid? A Skeptical Response to the Judicial Salary Debate. *The Journal of Legal Analysis*, 1, 1 2009).

[29] Schwarz, N. and Strack, F. Context Effects in Attitude Surveys: Applying Cognitive Theory to Social Research. *European Review of Social Psychology*(Jan 1 1991).

[30] Herr, P. M., Sherman, S. J. and Fazio, R. H. On the consequences of priming: assimilation and contrast effects. *Journal of experimental social psychology(Print)*(Jan 1 1983).

[31] Bavelas, A. Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America*(Jan 1 1950).

[32] Nosofsky, R. M. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*(Jan 1 1986).

[33] Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J. and Couper, M. Psychological Research Online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *Am. Psychol.*, 59, 2 2004), 105-117.