Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing

Andrew Mao
Harvard University
mao@seas.harvard.edu

Ece Kamar
Microsoft Research
eckamar@microsoft.com

Yiling Chen
Harvard University
yiling@eecs.harvard.edu

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

Megan E. Schwamb
ASIAA
Yale Center for Astronomy & Astrophysics
mschwamb@asiaa.sinica.edu.tw

Chris J. Lintott
University of Oxford
Adler Planetarium
cjl@astro.ox.ac.uk

Arfon M. Smith
Adler Planetarium
arfon@zooniverse.org

Abstract

Paid and volunteer crowd work have emerged as a means for harnessing human intelligence for performing diverse tasks. However, little is known about the relative performance of volunteer versus paid crowd work, and how financial incentives influence the quality and efficiency of output. We study the performance of volunteers as well as workers paid with different monetary schemes on a difficult real-world crowdsourcing task. We observe that performance by unpaid and paid workers can be compared in carefully designed tasks, that financial incentives can be used to trade quality for speed, and that the compensation system on Amazon Mechanical Turk creates particular indirect incentives for workers. Our methodology and results have implications for the ideal choice of financial incentives and motivates further study on how monetary incentives influence worker behavior in crowdsourcing.

Over the last decade, crowdsourcing has emerged as an efficient way to harness human intelligence for solving a wide range of tasks. While some crowdsourcing is unstructured and organic, such as efforts to coalesce knowledge on topics in Wikipedia and software applications created by open source projects, several crowdsourcing systems provide a structured environment that connects participants or *workers* with microtasks that are well-defined and self-contained. These systems typically do not require workers to be experts or to have strong familiarity with a task before starting to contribute. We shall use the terms crowdsourcing and crowd work interchangeably to refer to work done in these types of systems.

In paid crowd work, workers are compensated for completing tasks created by *requesters* in a marketplace or other assignment mechanism. Online marketplaces for specifying tasks and recruiting crowd workers include Amazon Mechanical Turk (MTurk), oDesk, and CrowdFlower. MTurk hosts a large variety of tasks, including data verification, language translation, and audio transcription. Other tasks include studies of human computation techniques and behavioral experiments (Ipeirotis 2010). Workers performing tasks through MTurk are often aware of their compensation and self-organize to find the best-paying and most interesting tasks (Chandler, Mueller, and Paolacci 2013).

Although crowdsourcing in the absence of monetary incentives has appeared in many forms, volunteer-based crowd work has recently expanded to organized platforms. Perhaps one of the most well known of these is the *Zooniverse*,[1] an online *citizen science* platform that connects scientists seeking human eyes on large amounts of data with participants (workers) interested in contributing to science (often called *citizen scientists*), and has been successful in producing valuable data for research. Examples of Zooniverse projects include Galaxy Zoo (Lintott et al. 2008), where galaxies are classified according to their shapes, and Planet Hunters[2] (Fischer et al. 2012; Schwamb et al. 2012; Lintott et al. 2013), where participants identify potential signals of planets orbiting distant stars. Citizen science systems rely solely on voluntary contributions of amateur participants without providing any monetary compensation, and volunteers run the gamut from a core community with strong intrinsic motivation (e.g. interest in a scientific discipline) to casual participants who visit the site once and leave (Raddick et al. 2013). Volunteers in unpaid crowdsourcing systems are driven by different motivations than workers of paid crowdsourcing platforms; volunteer crowd workers seek different objectives, and some may be more knowledgeable about a specific task than most workers in paid systems.

The many differences in motivation and incentives between paid and unpaid crowd work are not yet well understood, and a primary question is to characterize how different types of financial incentives influence the behavior of paid workers relative to volunteers. These differences are especially interesting with regard to the influence of incentives on the performing of tasks that are ambiguous or difficult, as different financial incentives may influence the amount of time workers spend on tasks and the quality of work performed. If workers are motivated solely by monetary compensation on a platform with no quality control, economic theory predicts that they will shirk and produce work of minimally acceptable quality. For example, the method currently used in paid crowdsourcing markets is to pay for each task, and this may naturally cause workers to complete tasks as fast as possible at the potential expense of accuracy. Even if workers exert a good-faith effort, the method of payment

[1] http://www.zooniverse.org
[2] http://www.planethunters.org

may still influence their work, as the requester and even workers themselves may not be explicitly aware of the way their work is influenced by financial incentives.

In this work, we adapt an annotation task originally performed by volunteers in the Planet Hunters citizen science project to an experiment with paid crowd workers on MTurk. With this experiment, we aim to answer the following questions:

- *How does the performance of workers in paid crowdsourcing environments compare to that of volunteers in unpaid crowdsourcing?*
- *What differences are produced in terms of accuracy, types of errors, speed, and engagement by different financial incentives for workers being paid on a task?*

In a set of experiments, we observe workers completing a variable, self-determined number of tasks under one of three different financial payment schemes. While the actual payments in our experiments do not depend on the quality of work produced, we use a gold standard to evaluate the quality and accuracy of work produced. Because workers select the number of tasks to complete and how quickly to work, we can measure the effect of payments on speed and worker engagement in terms of total time spent and the number of tasks completed. Our results do not provide a universal answer to the questions we asked above for tasks of all types. However, we identify trends for the task that we study, and believe that the approach we use can be harnessed in the study of questions about the influence of requests and incentives on other tasks. Specifically, we find that

- With proper incentives, paid crowd workers can achieve comparable accuracy to volunteers working on the same task, and perhaps even work at a faster rate.
- Different payment schemes, while paying workers approximately the same amount, lead to significant differences in the quality of work produced and amount of time spent. Our results suggest that financial incentives can be used to control tradeoffs among accuracy, speed, and total effort within a fixed budget.

In addition to observations on worker accuracy and speed, the experiments provide insights about workers' cognitive investment on paid crowdsourcing tasks. In particular, workers' self-reports on reasons for quitting tasks bring into view aspects of the meta-environment of MTurk. We find via self-reports that a significant percentage of workers stop because they are concerned about the quality of their work—a notable contrast to the belief that workers are motivated purely by immediate monetary gains within paid markets. Overall, our results highlight the complex nature of paid crowdsourcing marketplaces and underscore the need for richer models of the relationships between incentives and worker behavior.

Related Work

In the context of paid crowdsourcing, researchers have studied how the magnitude of financial incentives affects work produced. Horton and Chilton (2010) conducted an experiment to estimate the reservation wage of workers in MTurk. Mason and Watts (2009) examined financial rewards for two tasks, where workers were paid a fixed payment for each task completed and had the option of continuing to work on more tasks. They found that workers completed more tasks for a higher fixed payment, but that quality did not improve. Rogstadius et al. (2011) made a similar observation in their experiments. Yin, Chen, and Sun (2013) found that, while the magnitude of performance-contingent payments alone did not influence the quality of work produced, the change in the payment level for tasks in the same session did—increasing and decreasing payments increased and decreased the quality of work, respectively. Harris (2011) studied performance-contingent financial incentives (both rewards and penalties) and showed that the quality of work was higher in the presence of such incentives than in their absence. A large literature in economics and social psychology explores the relationships between the magnitude of financial compensation and productivity. We refer interested readers to a comprehensive review and meta-analysis by Camerer and Hogarth (1999).

Less attention has been focused on the influence of different payment schemes on the quality and quantity of work produced. Mason and Watts (2009) experimentally compared piece-rate schemes, where workers are paid for each task, and quota-based payment schemes, where workers are paid only after completing a bundle of tasks. They found that the quota-based scheme elicited higher effort from workers, while workers completed fewer tasks under the piece-rate scheme. Shaw, Horton, and Chen (2011) compared 14 financial, social, and hybrid incentive schemes, including performance-contingent reward and penalty, in their MTurk experiments. They identified two schemes where higher-quality work is produced in situations where workers' payments depend on the responses of her peers. Prior work in economics explored the influence of providing piece-wise payments versus an hourly wage. In a comprehensive study of the Safelite Glass Corporation (Lazear 2000), where workers install glass windshields in automobiles, a switch from hourly wages to piece-rate pay resulted in the firm becoming 44% more productive and workers earning higher wages overall. These results were obtained under an intrinsic policy that discouraged the temptation to do low-quality piece-rate work.

The motivation of volunteers in citizen science projects is much less studied; see Raddick et al. (2013) for one recent exception. Our work moves beyond the prior literature in several ways. First, we compare volunteer and paid workers on the same task. Second, we focus on the influence of different payment schemes within a comparable budget. Third, we provide evidence of secondary meta-incentives in paid crowdsourcing, using MTurk as an example.

Task Model

We consider a challenging citizen science task that, by its very nature, invites a high degree of variability in annotations. The task is analogous to finding needles in a sequence of haystacks. Each task can viewed as a haystack housing needles of interest. Workers examine the data and can mark needles directly, but the task is ambiguous because workers may miss certain needles or falsely mark other regions

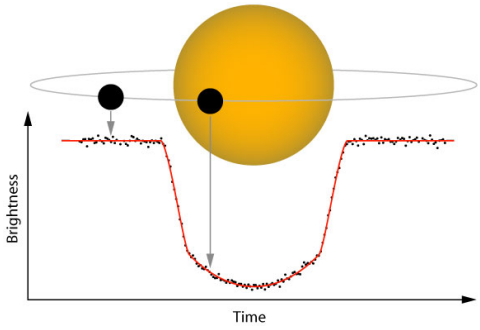


Figure 1: A light curve, showing the transit method of detecting exoplanets.

depending on varying levels of difficulty. By exerting more effort in a more detailed investigation, workers can generally obtain higher accuracy on this task. Workers may complete several tasks in sequence in a continuous session.

Many human computation tasks fall into this category, such as annotating events of interest in images (Salek, Bachrach, and Key 2013). We find such task domains particularly interesting because the worker's perception of the truth can be ambiguous: workers can produce both *false positives* when regions are wrongly marked and *false negatives* when objects of interest are missed, even if they are doing their best. Hence, the particular financial incentives at hand may influence the worker's contribution and amplify the types of errors the worker makes. We now turn to the specifics of the haystacks and needles that we have studied.

Planet Hunters

Planet Hunters (Schwamb et al. 2012) is a citizen science project started in December 2010 with the goal of finding planets orbiting around distant stars (extrasolar planets or exoplanets), where volunteers search for the signatures of exoplanets in the data from the Kepler spacecraft (Borucki et al. 2010).

The Kepler spacecraft is a space-based telescope that simultaneously monitors the brightness of over 160,000 stars, producing graphs called *light curves* for each star. Kepler generates two data points per hour by measuring the brightness of a star approximately every 30 minutes. A planet that is orbiting the star in a plane aligned with the viewing angle of the telescope will partially obscure the star once per orbit in a *transit*, causing the observed brightness to drop and corresponding dip in the light curve (see Figure 1; (Winn 2010 and references within). Typical transits last from two to dozens of hours, so the telescope records multiple data points for a typical transit. The size of the dip in the light curve is proportional to the surface area of the star and the planet; the *relative transit depth*, or percentage decrease in the brightness of a star obscured by a planet during a transit, can be computed from the radius of the planet R_p and the star R_*:

$$\text{relative transit depth} = \frac{R_p^2}{R_*^2}. \quad (1)$$

For example, to a distant observer, Jupiter would obscure the sun by around 1%, while the Earth obscures only 0.01%.

Several aspects of the transit detection task affect its difficulty. Telescopes have a natural instrumentation error when measuring the brightness of a star. Moreover, the brightness of stars themselves vary over time, causing fluctuations and changes in the light curve (typically on timescales longer than transits). A transit with a small relative transit depth can be easily seen in a low-variability light curve while a transit with a large relative transit depth may be even hard to see in a highly variable light curve. A planet with short period (orbit time) compared to the span of observation can cause multiple transits to appear at regular intervals, making detection easier. In general, transits by fast-moving planets, by small planets, and in front of large stars are more difficult to detect.

Although the transit method has been in use by astronomers, the orbital telescope technology deployed in Kepler has allowed for searches of planets en masse. Planet Hunters enlists human volunteers to review Kepler data, and has resulted in several planet discoveries that were not detected by automated methods, demonstrating the value of human pattern recognition for this task (Fischer et al. 2012; Lintott et al. 2013).

Experiment Design

Interface The interface for Planet Hunters is open-ended, allowing workers to freely examine a light curve of 1,600 data points collected through ∼35 days with tools for zooming and drawing simple boxes around any potential transits that they see. We designed a similar interface for an MTurk task, shown in Figure 2. Workers can mark possible transits on a light curve by drawing a box, resizing and/or deleting them as desired. In this way, workers produce annotations in a continuous space, defined by the coordinates and size of the boxes.

After accepting our HIT and reading a short consent form, workers see an interactive tutorial of the interface, describing what a light curve is and how planet transits are detected. The tutorial demonstrates zoom controls and the annotation process. A help menu is available at any point during the task that provides additional information about identifying planet transits and the interface controls.

A key aspect of this experiment is that workers can annotate multiple light curves (finish multiple tasks), choosing the amount of effort they want to contribute before quitting. This is similar to the process of performing a number of tasks for a particular requester on systems like MTurk. At each light curve, workers may choose to continue on to the next light curve or to finish their work and submit the task. We place some basic controls on the experiment such as limits on participation, described later in this section.

When workers choose to complete their work, they are required to complete a short survey containing questions about the HIT. We ask workers whether they think the HIT was easy, fun, or well paid, about their strategy on the task, and if they ran into any bugs. Most importantly, we ask workers why they decided to stop working and to submit the task and what, if anything, would have made them work longer.

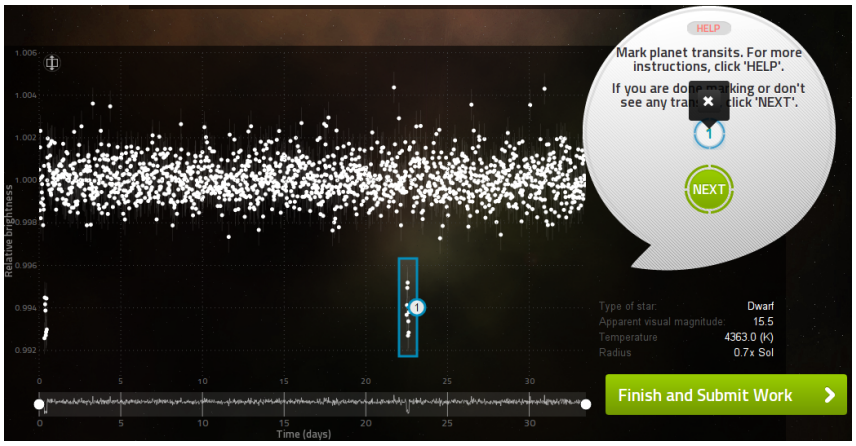


Figure 2: The experiment interface, showing a single annotated transit. The **Next** button accesses the next light curve, while the **Finish** button submits the task.

Simulated Transits In contrast to most real-world crowdsourcing tasks, transit detection has the useful feature that realistic data with a ground truth can be generated easily; simulated transits of different depths and durations can be added to an observed light curve. The Planet Hunters team injected simulated transits into real Kepler data (Schwamb et al. 2012) to estimate the efficiency of detecting different types of planet transits using crowdsourcing.

While planets with multiple visible transits should be easier to detect in a light curve, Schwamb et al. (2012) showed that the difference in behavior in Planet Hunters is insignificant for orbital periods less than 15 days, and that transit depth is the dominant effect. Therefore, we define a difficulty measure for detecting transits for a simulated light curve by comparing the relative transit depth (Equation 1) to the noise of the light curve:

$$\text{difficulty} = \frac{\text{stdev(differences in adjacent points)}}{\text{relative transit depth}} \qquad (2)$$

A light curve simulation with difficulty 0.2 means that the depth of the transit will be 5 times bigger than the typical noise in the graph, and should be relatively easy to spot. A simulation with difficulty 1 means that a transit can easily hide within the noise in the graph, and is more easy to spot. Simulations of difficulties greater than 1 should be very difficult to detect. In the experiment described in the following section, we use simulated light curves that had been annotated earlier by volunteers on the Planet Hunters project.

Payment Schemes The primary goal of the experiment is to compare the effects of different payment schemes on workers' performance on an ambiguous task at various levels of difficulty, and to performance on the same task by volunteers. We consider three different non-performance-contingent payment schemes:

- **Pay per task.** Workers are paid for each task that they complete; in our case, this is per light curve. This is a typical payment scheme for paid microtasks.
- **Pay for time.** Workers are paid for each unit time that they spend working, regardless of their actual output. This payment scheme is employed commonly in traditional employment.
- **Pay per annotation.** Workers are paid for each object that they annotate; in our case, this is per marked transit.

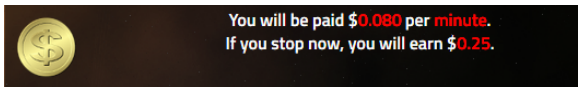


Figure 3: An sample payment message displayed to workers.

We focus on payment schemes that do not depend on the accuracy or quality of workers' output. These payment schemes are simple to explain to workers and do not require the implementation of a quality control mechanism. To ensure that workers are fully aware of how they are getting paid, we show a continually updated banner at the top of the task (Figure 3) which displays the method of payment and how much they have earned so far.

Data Selection and Treatments To allow for comparison between unpaid volunteers and paid workers, we selected light curves for our experiment from the set of light curves that had been already reviewed by numerous volunteer citizen scientists contributing to Planet Hunters. All of the light curves were collected during Quarter 1 of the Kepler Mission. As transits are rarely seen overall, our dataset must include many light curves with no transits so that the task is realistic and workers don't expect to see them in every light curve. However, we also need to have sufficient simulations to obtain data about the accuracy of workers. Based on the original annotation results, we removed pathological light curves from the data that were particularly confusing, including planets periods shorter than 5 days. We ultimately selected a set of 250 light curves with up to 6 simulated transits and an additional 750 light curves without simulated transits (devoid of planet transits to the best of our knowledge). The simulated light curves are distributed approximately uniformly in the difficulty measure described

in Equation 2, from values ranging from 0.2 to 1.0. We chose this range after visually inspecting many light curves, as the range included examples that were neither too obvious nor too difficult to detect.

We adopt notions of *precision* and *recall* from the information retrieval community to measure the accuracy of the annotations. An annotated box is counted as correctly marked if the center of the box is within an actual transit—this simple measure of correctness is convenient because it allows for some latitude in the width of a box drawn, which in turn depends on the zoom level. The precision of a worker's annotations is the fraction of his annotations that are actual transits, or the ratio of the number of correct annotations to the total number of annotations. The recall of a worker's annotations is the fraction of transits that are annotated by the worker, or the ratio of the number of transits correctly annotated to the total number of transits.

Controls and Monitoring Paying workers without regard to quality can lead to low quality output should workers behave as purely economic agents and expect no negative consequences for errors. As a result, we create controls that would be expected in a practical implementation.

- **Minimum of 5 seconds per light curve.** Without this control, a worker being paid by task can potentially click through light curves very quickly and be fully paid for almost no work.
- **Maximum of 8 annotations per light curve.** In absence of this control, a worker being paid per annotation may mark a potentially infinite number of false positives and be paid for each one. We restrict the minimum orbital period in our data to be > 5 days, so at most 6 transits will appear.
- **Maximum of 3 minutes of inactivity.** Without this contol, a worker being paid by time can potentially do nothing while earning wages. An inactivity warning is shown when a worker has done nothing for 2 minutes. If the worker continues to do nothing for a total of 3 minutes, the task ends and automatically redirects the worker to the exit survey.

We record all of the above events during a session. By monitoring inactivity and enforcing a timeout on the task, we are able to detect when a worker is no longer paying attention or has become distracted. As workers must complete the exit survey to submit the HIT, we can learn why they stopped working. We also restrict all worker sessions to a maximum of one hour or 200 light curves, to limit the amount of data from any one particular worker.

In addition to the detection of timeout, we track the amount of inactivity for each worker during their session, defined by the total amount of time that they were inactive for 30 seconds or more.

Hypotheses When worker payments do not depend on performance, workers would theoretically behave in extreme ways to maximize short-term payment. In theory, workers being paid by annotation would mark as many transits as possible (mostly incorrectly), earning the fixed amount for each. Workers being paid by task would click through the light curves very quickly, paying minimal attention to each one. And workers being paid by time might be expected to simply sit through a task and do barely anything, earning their hourly wage without spending much effort. However, we would not expect to see these extreme behaviors in practice. Workers typically expect that they will be evaluated in some way for their work, and many MTurk workers are keenly aware that rejected work will prevent them from doing lucrative tasks in the future. Aside from spammers, most workers will try to follow the instructions and do the task as well as they can understand it. Yet, the ambiguous nature of the task of identifying planets means that workers cannot be completely sure about the 'wrong' or 'right' answers; a worker being paid by annotation may subconsciously "see" more transits than a worker being paid by task, without being overtly dishonest. How strong might this psychological bias be?

The difficulty level of the task may also affect workers' accuracy. When transits are plainly obvious in a light curve, we might expect all but the laziest workers to mark them. However, when transits are more ambiguous, we might expect workers who are paid per light curve or by time to more likely overlook them.

Most interestingly, the demographics of volunteer and paid workers are very different. Workers on Planet Hunters consist of many one-time users, but also include a dedicated community of users with an active discussion forum and many very motivated amateur astronomers, combing the data for transits and even writing their own analysis code. On the other hand, MTurk workers in our experiment do this task with nothing but a short tutorial, and are given a payment in return for their efforts. Given the differences in background and motivation, which group will do better?

Limitations of Comparison The focus of our experiment is to compare payment schemes, but we also give a comparison to volunteer work. There are some notable differences between our experiment and the original interface used by volunteers (see Schwamb et al. 2012 for a full description), which presents a series of additional questions to registered users. Our interface focuses only on transit annotation, and uses a free-drawing interaction and a different tutorial geared toward MTurk workers. Schwamb et al. show consistent behavior between the original box-placement annotation method used in Planet Hunters and a free-drawing method using a similar performance metric, but our measure of accuracy is more strict.

Results

We conducted our experiment as a between-subject study where all workers were allowed to do the HIT exactly once, to reduce the effect of noise in the results from worker experience over repeated tasks. In each set of experiments, workers were randomly assigned to one of the payment treatments. Workers were assigned a new, randomly selected light-curve each time they continued in the task.

Treatment	N	Wage	Secs/Task	Anno/Task
volunteer	*	*	50*	1.250
$0.0453/annot.	71	$10.993	29.13	1.964
$0.0557/task	74	$8.056	24.89	1.435
$0.08/minute	71	$4.800	27.45	1.454

Table 1: Volunteer and worker behavior in the pilot. N: number of experiment sessions; Wage: average hourly wage; Secs/Task: average number of seconds per task; Anno/Task: average number of annotations labeled per light curve. *In this table and Table 2, volunteers may work for longer due to possible additional questions in the task; we also omit statistics that would be misleading given the differences described previously.

Initial Observations

To make meaningful comparisons among the treatments the wage across treatments must be comparable. Identifying comparable wages across schemes is tricky as we do not *a priori* know how workers would behave. Hence, we conducted a pilot experiment to observe the behavior of workers and obtain a better idea of what comparable wages would be.

Through our experience with tasks on MTurk and guidelines posted on various discussion forums, we observed that most experienced workers aimed at a target of $0.10/minute or $6.00/hr as a fairly reimbursed task for which they would continue to work indefinitely. We picked a lower wage of $4.80, which is close to a fair payment for worker time but low enough that we could expect workers to quit our task (before the time limit) and thus obtain information about why they left.

To set wages for the various treatments, we examined the behavior of unpaid citizen scientists on the corresponding subset of the existing Planet Hunters data, obtaining a baseline of how many annotations volunteers would mark and the rate at which they completed the light curves. Using this data, we computed a wage of $0.0557 per task and $0.0453 per annotation, which would all pay the same wage of $4.80 if the paid workers behaved similarly as the volunteers.

Table 1 shows a summary of observations from the pilot experiment. In the treatments shown, over 200 unique workers annotated about 14,000 light curves. Notably, paid workers completed tasks significantly more quickly than the volunteer workers, resulting in a much higher wage for both the task and annotation treatments. Moreover, workers in the annotation treatment were much more eager about marking transits than the other workers, showing a clear bias. This further boosted their wage to an average close to $11/hour; some workers were able to earn over $30/hour, and we observed many comments on various worker forums that our task paid extremely well.

The non-uniform effective hourly wage earned across the treatments confirms that paid workers behave significantly differently from volunteers, both working faster and being influenced by their financial incentives significantly. However, the large discrepancy between wages makes it difficult to compare the payment methods, as some workers are earning more than twice as much as others. We also observed some notable meta-effects during this experiment. As we monitored worker discussion forums over the course of several days, we noticed that workers had begun to discuss our task and compared their payments with each other, being especially curious as to why some thought the task was particularly well-paid compared to others. On the site where the discussion was most lively (http://www.mturkforum.com), we talked to workers and discovered that while there was actually a policy against discussing research studies, our task actually appeared to be a normal MTurk task (as we had intended apart from the consent process), and the normal appearance had prompted the discussion. We were pleasantly surprised to learn that the Turker community had self-imposed rules to protect the integrity of research, and were advised to include an explicit statement not to discuss the task with others so as to be covered by this policy.

Balanced Payments

The observations on the pilot study prompted us to design a second round of experiments where workers are paid more equally, and to eliminate biases caused by external discussion. For example, workers might produce worse quality work if they expected a certain level of payment in the task from discussion but received a much lower amount.

We made the assumption, based on aforementioned work in financial incentives, that the per-task behavior of workers would not change much compared to their payment level. Hence, we could scale the piece-rate wages for the annotation and task treatments accordingly, and obtain data where the effective hourly wage is closer to the target of $4.80. While we could not enforce this in advance, the treatments would be comparable as long as they resulted in similar levels of payment. This resulted in piece-rate payments of $0.0331 per light curve and $0.0197 per annotation.

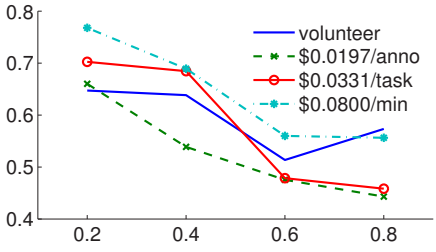
Second, we took several precautions to minimize external discussion about our task. We followed the advice of showing an explicit message not to discuss their work with others during the exit survey. We also posted on several discussion forums that participants should not discuss the task; we noticed that workers indeed passed on this message when others asked about this task. Moreover, since we required all workers to be unique, workers from the first set of experiments were not able to do the task, and this caused the amount of discussion to die down significantly. When closely monitoring discussions, we saw very few posts about our task during the experiment; workers also trickled in at a much slower rate compared to a veritable 'flood' of workers looking for a well-paid HIT in the first experiment.

Table 2 shows a summary of the second experiment. A total of 356 workers annotated over 17,000 light curves. Of particular note is that the effective hourly wage earned by workers was much closer together than in the previous treatment; the workers in the pay by annotation and by time treatments earned almost exactly the same amount, and the workers in the pay by task treatment, earned only slightly more.

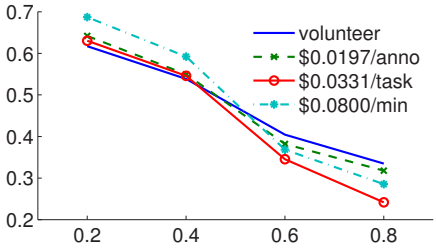
Accuracy by Difficulty. We split the set of simulated light curves into four buckets determined by the difficulty mea-

Treatment	N	Tasks	Wage	Tasks/Sess.	Secs/Task	Median Time	Anno/Task	Precision	Recall	Pct. Inactive
volunteer	*	*	*	*	50*	*	1.250	0.656	0.518	*
$0.0197/annot.	118	4629	$4.802	39.22	28.02	9:04	1.897	0.635	0.485	0.101
$0.0331/task	121	7334	$5.580	60.61	21.35	15:08	1.384	0.660	0.454	0.097
$0.08/minute	117	5365	$4.800	45.85	34.65	18:05	1.348	0.713	0.497	0.149

Table 2: Volunteer and worker behavior in the second experiment. N: number of experiment sessions; Task: total number of tasks completed in all experiment sessions; Wage: average hourly wage; Tasks/Sess.: average number of tasks completed per session; Secs/Task: average number of seconds spent on a task; Median Time: the median of the total time spent on an experiment session; Anno/Task: average number of annotations labeled for a light curve; Pct. Inactive: average percentage of time that a worker is detected inactive.



(a) Precision.



(b) Recall.

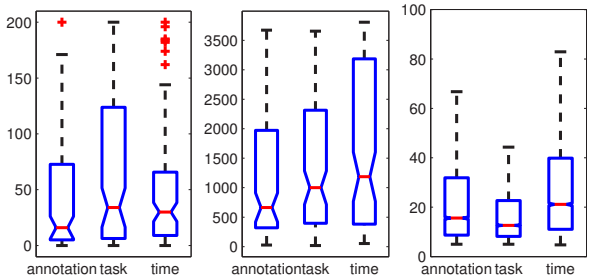
Figure 4: Accuracy by difficulty.

sure in Equation 2, and computed precision and recall for each bucket, displayed in Figure 4. As expected, both precision and recall drop at higher levels of difficulty, with the only exception being the volunteer group at the hardest difficulty bucket. To test the significance of differences between each bucket, we used a two-sided paired t-test between the aggregate false positive rate and false negative rate among the light curves in each bucket.

We make several notable observations from the second experiment. With regard to precision, paying by time leads to significantly higher performance than paying by annotation at all levels of difficulty (for all but the last bucket, $p < 0.005$). Paying by annotation shows by far the worst precision across the board, with many differences being highly significant. We note that the precision across the volunteer population decreases more slowly as difficulty increases: at the easiest difficulty, they show significantly worse precision than the task and time treatments. However, for the most difficult tasks, they show significantly better precision than for the task and annotation treatments. We discuss possible reasons for this below.

For recall, workers paid by time show by far the best recall for easy tasks. However, the volunteers and workers paid by annotation show best recall at high levels of difficulty. Workers paid by task generally perform poorly, and in the most difficult bucket, they show the worst recall by far ($p < 0.002$ compared to the unpaid and annotation treatments). Similar to the observation made in the precision analysis, overall, we observe that the recall scores of the volunteers are less sensitive to the difficulty level than the paid workers.

Worker Attention. We can measure the attention or interest of workers in two ways: by the amount of time they are spending on each task, a measure we believe roughly corresponds to effort; and the total amount of time in the session. This comparison is particularly interesting because workers are being paid roughly the same amount for their time, with the wage being almost identical for the annotation and time treatments. Table 2 shows that the financial incentive scheme implemented has significant influences on the speed of workers for completing tasks. When being paid by time, workers spend over 60% more time on each task than when being paid for each task, and this is accompanied by a corresponding increase in accuracy. These findings suggest that payment methods can be used to trade off speed and accuracy in worker output. In addition, workers spend less time on the task and show significantly worse precision when paid by annotation rather than by time, in spite of earning almost the same hourly wage. Figure 5 shows the distribution of statistics for sessions. The difference in number of tasks per session is significant for workers paid by task compared to the other two treatments at the 0.05 level. The difference in the total time spent in a session is also significant at the 0.05 level for the time versus annotation treatments. The differences in seconds per task is highly significant ($p < 0.0001$) for all treatments.



(a) # Tasks Done (b) Total Time (c) Seconds/Task

Figure 5: Distribution of session statistics. Boxplots show top and bottom quartiles and median as red line.

We also examine the reasons that workers gave for ending the task in the exit survey. There are many explanations for exiting a task. Horton and Chilton (2010) suggested that workers may set a target earnings level when deciding when to stop. In our experiment, workers could be interrupted or time out. Using the experiment controls as well as workers' stated reasons for stopping, we classified reasons for stopping into different categories, described with examples as follows:

- **quality** – concerned about submitting bad work or being rejected: "*I decided to stop because I wasn't sure if I was doing a good job or not. I would have continued, but I did not want my HIT to be rejected because I misunderstood or provided bad data.* "
- **limit** – reached a limit of 200 tasks or one hour.
- **exogenous** – had another prior commitment that had to be completed. Surprisingly, some employees Turk during their regular jobs: "*I had to go back to work...I would have worked longer if my lunch break was longer.* "
- **interruption** – temporarily interrupted during the task, but intended to return to it. This included many bathroom breaks, phone calls, and pizza deliverymen arriving.
- **pay** – The pay for the task was too low.
- **bored / tired** – bored or tired of the task.
- **technical** – didn't seem to understand the task or had a technical problem.
- **target** – reached a target self-imposed time or monetary amount: "*I decided that $2.00 was enough for a single task and the amount of time spent on it. If I was paid much better I would have continued a bit longer; but I don't like doing a single task/hit for too long.* "

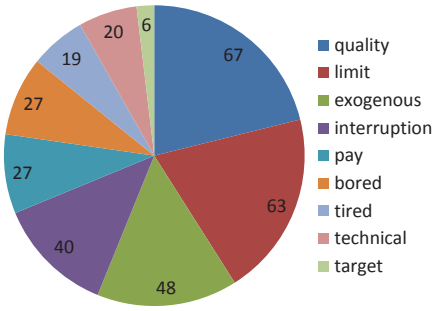


Figure 6: Classification of reasons for quitting.

Figure 6 shows that many workers were interrupted by distractions or outside commitments or reached our limit. Surprisingly, a significant proportion of workers chose to stop because they were unsure of the quality of their work. This runs counter to characterizations of Turkers as greedy workers who maximize their short-term rewards. To understand this phenomenon further, we analyzed workers' comments carefully and communicated with them on discussion forums. It became clear that this behavior was founded in two goals. First, workers did not want to have their work rejected, which would waste their effort and lower their HIT approval rate (used as a filter on many tasks). Therefore, if workers are more uncertain about a requester's approval policy, they would do less work to 'test the water'. Second, some workers were actually concerned about providing good quality work to requesters and submitted our hit early or even returned it when they were unsure about their work. Very few workers explicitly mentioned a payment or time goal as a reason for stopping the task. As mentioned in Chandler, Mueller, and Paolacci (2013), it is very important for researchers to be aware of these meta-incentives when designing tasks and especially experiments for paid workers.

Discussion

To our knowledge, our experiments, centering on challenging, ambiguous annotation tasks of varying levels of difficulty, provide the first comparison of volunteer workers to workers paid by different financial schemes in an online task market. Under the tasks we studied, we find comparable performances between volunteers and appropriately paid workers. We note that the results obtained via experiments with the planet discovery task may not generalize to other tasks. However, the overall approach and methodology can provide the basis for analogous studies. Also, the results have general implications on strategies for compensating workers in online task markets. We found that worker behavior is sensitive to variation of methods of payment. We believe that such influences of payment scheme on worker behavior is a feature rather than a drawback: paying workers the same effective wage, but with different piece-rate methods, can be used to trade off precision, recall, speed, and total attention on tasks. In our case, the canonical per-task payment used on MTurk and many other task markets results in the fastest task completion, but lowest recall. Other methods of payment, such as paying a wage, caused workers to work more slowly, but with better results. Being able to selectively control the output of human workers is desirable for many algorithms that use human computation, and the use of financial incentives in this way is an effective lever that warrants further careful study.

We also observed that the payment methods vary in their sensitivity to difficulty level, and this finding suggests that the performance of volunteers and workers paid using different methods may vary in sensitivity to the hardness of the task. For the planet discovery task, workers being paid in the canonical per-task scheme showed the greatest drop in precision as difficulty increased. The findings suggest that the design of financial incentives is important in achieving a desired level of performance from crowd workers for a heterogeneous set of tasks. We believe that we have only scratched the surface in exploring the differences in incentives between unpaid citizen science projects and paid crowdsourcing platforms. Comparing the motivations of workers in each of these settings is an important problem that warrants further study.

Our experiments indicate that, even in paid task markets, indirect or secondary incentives may influence the behavior of workers. When examining the reasons that microtask workers may get distracted or leave, we find that many workers report being concerned about the quality of their work.

On the other hand, it is likely that some of these workers may behave differently and provide low-quality work in response to a task with loose controls or to a requester with low standards. However, this also suggests that, over the short term and with the right controls, one can indeed use different non performance-contingent payment schemes to collect high quality data from workers. Overall, our collective observations highlight multiple opportunities and directions with pursuing deeper understanding of how incentives influence the behavior and output of crowd workers.

Acknowledgements This research was commenced during an internship at Microsoft Research and was later partially supported by the NSF under grant CCF-1301976. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone. Schwamb was supported in part by a NSF Astronomy & Astrophysics Postdoctoral Fellowship under award AST-1003258. We thank Matt Giguere and Debra Fischer for creating the simulated planet transit data, and John Johnson for providing Figure 1.

References

Borucki, W. J.; Koch, D.; Basri, G.; Batalha, N.; Brown, T.; Caldwell, D.; Caldwell, J.; Christensen-Dalsgaard, J.; Cochran, W. D.; DeVore, E.; Dunham, E. W.; Dupree, A. K.; Gautier, T. N.; Geary, J. C.; Gilliland, R.; Gould, A.; Howell, S. B.; Jenkins, J. M.; Kondo, Y.; Latham, D. W.; Marcy, G. W.; Meibom, S.; Kjeldsen, H.; Lissauer, J. J.; Monet, D. G.; Morrison, D.; Sasselov, D.; Tarter, J.; Boss, A.; Brownlee, D.; Owen, T.; Buzasi, D.; Charbonneau, D.; Doyle, L.; Fortney, J.; Ford, E. B.; Holman, M. J.; Seager, S.; Steffen, J. H.; Welsh, W. F.; Rowe, J.; Anderson, H.; Buchhave, L.; Ciardi, D.; Walkowicz, L.; Sherry, W.; Horch, E.; Isaacson, H.; Everett, M. E.; Fischer, D.; Torres, G.; Johnson, J. A.; Endl, M.; MacQueen, P.; Bryson, S. T.; Dotson, J.; Haas, M.; Kolodziejczak, J.; Van Cleve, J.; Chandrasekaran, H.; Twicken, J. D.; Quintana, E. V.; Clarke, B. D.; Allen, C.; Li, J.; Wu, H.; Tenenbaum, P.; Verner, E.; Bruhweiler, F.; Barnes, J.; and Prsa, A. 2010. Kepler Planet-Detection Mission: Introduction and First Results. *Science* 327:977–.

Camerer, C. F., and Hogarth, R. M. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* 19(1-3):7–42.

Chandler, J.; Mueller, P.; and Paolacci, G. 2013. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*.

Fischer, D. A.; Schwamb, M. E.; Schawinski, K.; Lintott, C.; Brewer, J.; Giguere, M.; Lynn, S.; Parrish, M.; Sartori, T.; Simpson, R.; Smith, A.; Spronck, J.; Batalha, N.; Rowe, J.; Jenkins, J.; Bryson, S.; Prsa, A.; Tenenbaum, P.; Crepp, J.; Morton, T.; Howard, A.; Beleu, M.; Kaplan, Z.; vanNispen, N.; Sharzer, C.; DeFouw, J.; Hajduk, A.; Neal, J. P.; Nemec, A.; Schuepbach, N.; and Zimmermann, V. 2012. Planet Hunters: the first two planet candidates identified by the public using the kepler public archive data. *Monthly Notices of the Royal Astronomical Society* 419(4):2900–2911.

Harris, C. 2011. You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, 15–18.

Horton, J. J., and Chilton, L. B. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce (EC)*, 209–218. New York, NY, USA: ACM.

Ipeirotis, P. G. 2010. Analyzing the amazon mechanical turk marketplace. *XRDS* 17(2):16–21.

Lazear, E. P. 2000. Performance pay and productivity. *The American Economic Review* 90(5):pp. 1346–1361.

Lintott, C. J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M. J.; Nichol, R. C.; Szalay, A.; Andreescu, D.; Murray, P.; and Vandenberg, J. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389:1179–1189.

Lintott, C. J.; Schwamb, M. E.; Barclay, T.; Sharzer, C.; Fischer, D. A.; Brewer, J.; Giguere, M.; Lynn, S.; Parrish, M.; Batalha, N.; Bryson, S.; Jenkins, J.; Ragozzine, D.; Rowe, J. F.; Schwainski, K.; Gagliano, R.; Gilardi, J.; Jek, K. J.; Pkknen, J.-P.; and Smits, T. 2013. Planet Hunters: New Kepler planet candidates from analysis of quarter 2. *The Astronomical Journal* 145(6):151.

Mason, W., and Watts, D. J. 2009. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, Proceedings of the 1st Human Computation Workshop (HCOMP), 77–85. New York, NY, USA: ACM.

Raddick, M. J.; Bracey, G.; Gay, P. L.; Lintott, C. J.; Cardamone, C.; Murray, P.; Schawinski, K.; Szalay, A. S.; and Vandenberg, J. 2013. Galaxy Zoo: Motivations of citizen scientists. *Astronomy Education Review* 12(1):010106.

Rogstadius, J.; Kostakos, V.; Kittur, A.; Smus, B.; Laredo, J.; and Vukovic, M. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Salek, M.; Bachrach, Y.; and Key, P. 2013. Hotspotting – a probabilistic graphical model for image object localization. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI)*.

Schwamb, M. E.; Lintott, C. J.; Fischer, D. A.; Giguere, M. J.; Lynn, S.; Smith, A. M.; Brewer, J. M.; Parrish, M.; Schawinski, K.; and Simpson, R. J. 2012. Planet Hunters: Assessing the Kepler inventory of short-period planets. *The Astrophysical Journal* 754(2):129.

Shaw, A. D.; Horton, J. J.; and Chen, D. L. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, 275–284. New York, NY, USA: ACM.

Winn, J. N. 2010. Transits and occultations. Chapter of the graduate-level textbook, EXOPLANETS, ed. S. Seager, University of Arizona Press. http://arxiv.org/abs/1001.2010.

Yin, M.; Chen, Y.; and Sun, Y.-A. 2013. The effects of performance-contingent financial incentives in online labor markets. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, AAAI '13.