Running experiments on Amazon Mechanical Turk

Gabriele Paolacci*
Advanced School of Economics, Ca' Foscari University of Venice

Jesse Chandler
Woodrow Wilson School of Public and International Affairs, Princeton University

Panagiotis G. Ipeirotis
Leonard N. Stern School of Business, New York University

Abstract

Although Mechanical Turk has recently become popular among social scientists as a source of experimental data, doubts may linger about the quality of data provided by subjects recruited from online labor markets. We address these potential concerns by presenting new demographic data about the Mechanical Turk subject population, reviewing the strengths of Mechanical Turk relative to other online and offline methods of recruiting subjects, and comparing the magnitude of effects obtained using Mechanical Turk and traditional subject pools. We further discuss some additional benefits such as the possibility of longitudinal, cross cultural and prescreening designs, and offer some advice on how to best manage a common subject pool.

Keywords: experimentation, online research

1 Introduction

Mechanical Turk started in 2005 as a service to "crowdsource" labor intensive tasks and is now being used as a source of subjects for experimental research (e.g. Eriksson & Simpson, 2010; Alter et al., in press). However, a combination of unfamiliarity with what online labor markets are (and how to use them), uncertainty about the demographic characteristics of their participants and concerns about data quality from this sample may make some researchers wary of using Mechanical Turk to collect data. To address these concerns we report demographic characteristics of Mechanical Turk workers, highlight some of the unique practical and methodological strengths of Mechanical Turk as a source of research subjects and compare classic judgment and decision making effects in this population and more traditional subject populations.

The article is organized as follows. In Section 2, we introduce the main features of Mechanical Turk and demonstrate that the population of Mechanical Turk is at least as representative of the U.S. population as traditional subject pools. Further, we show that it is shifting to include more international participants. In Section 3, we review the logic underlying concerns with collecting data using Mechanical Turk and present the strengths and potentials of Mechanical Turk relative to other online and offline methods of recruiting subjects. In Section 4, we present the results of a comparative study involving classic experiments in judgment and decision-making; we found no differences in the magnitude of effects obtained using Mechanical Turk and using traditional subject pools. Section 5 concludes by offering some advice on payment of individual subjects.

2 Amazon Mechanical Turk

2.1 Mechanical Turk: The service

Mechanical Turk is a crowdsourcing web service that coordinates the supply and the demand of tasks that require human intelligence to complete. Mechanical Turk is named after an 18th century chess playing "automaton" that was in fact operated by a concealed person. It is an online labor market where employees (called *workers*) are recruited by employers (called *requesters*) for the execution of tasks (called *HIT*s, acronym for Human Intel-

*The authors thank Massimo Warglien and contributors to the Experimental Turk blog for useful discussions, and the editor Jonathan Baron for helpful suggestions. Gabriele Paolacci receives financial support by Fondazione Coin. Panagiotis G. Ipeirotis is supported by the National Science Foundation under Grant No. IIS-0643846 and by a New York University Research Challenge Fund. Address: Gabriele Paolacci, Advanced School of Economics, Ca' Foscari University, Cannaregio 873, 30121 Venice, Italy. Email: paolacci@unive.it.

ligence Tasks) in exchange for a wage (called a *reward*). Both workers and requesters are anonymous although responses by a unique worker can be linked through an ID provided by Amazon. Requesters post HITs that are visible only to workers who meet predefined criteria (e.g., country of residence or accuracy in previously completed tasks). When workers access the website, they find a list of tasks sortable according to various criteria, including size of the reward and maximum time allotted for the completion. Workers can read brief descriptions and see previews of the tasks before accepting to work on them.

Tasks are typically simple enough to require only a few minutes to be completed such as image tagging, audio transcriptions, and survey completion. More complicated tasks are typically decomposed into series of smaller tasks including the checking and validation of other workers' HITs. Once a worker has completed a task, the requester who supplied that task can pay him. Rewards can be as low as $0.01, and rarely exceed $1. Translated into an hourly wage, the typical worker is willing to work for about $1.40 an hour (Horton & Chilton, in press).

A requester can reward good work with bonuses and punish poor quality work by refusing payment or even blocking a worker from completing future tasks. Requesters who fail to provide sufficient justification for rejecting a HIT can be filtered out by a worker, preventing future exploitation. Some may wonder about who is willing to work for so low wages. With the goal of providing experimenters with a typology of the recruitable workforce in Mechanical Turk, we now present the results of a demographic survey conducted in February, 2010.

2.2 Demographics of Mechanical Turk

It is reasonable to assume that only those in poor countries would be willing to work for such low wages. However, until recently, Amazon.com was paying cash only to workers that had a bank account in the U.S., with other workers paid with Amazon.com gift cards. This policy discouraged workers from other countries, and past demographic surveys found that 70–80% of workers were from the United States and that Mechanical Turk workers were relatively representative of the population of U.S. Internet users (Ipeirotis, 2009; Ross et al., 2010). Recently, however, the population dynamics on Mechanical Turk have changed significantly, with a greater proportion of Indian subjects in recent experiments (e.g. Eriksson & Simpson, 2010), suggesting the need for a fresh survey of the workers.

We collected demographics of 1,000 Mechanical Turk users. The survey was conducted over a period of three weeks in February 2010. Each respondent was paid $0.10 for participating in the survey, which required 3 minutes on average to complete. This resulted in an hourly average wage of $1.66, which is superior to the median reservation wage of $1.38/hour (Horton & Chilton, in press). Participants from 66 countries responded. The plurality of workers was from the United States (47%), but with a significant number of workers from India (34%). We will now present the demographics for American workers. A more detailed breakdown of demographics (including tables and graphs and an analysis for India-based workers) is presented by Ipeirotis (2010).

Gender and age distribution. Across U.S.-based workers, there are significantly more females (64.85%) than males (35.15%). The relative overabundance of women is consistent with research on subjects recruited through the Internet (Gosling et al., 2004) and may reflect women having greater access to computers (either at home or at work) or gender differences in motivation. Workers who took part to our survey were 36.0 years old on average (min = 18, max = 81, median = 33) and thus slightly younger then both the U.S. population as a whole and the population of Internet users.

Education and income level. In general, the (self-reported) educational level of U.S. workers is higher than the general population. This is partially explained by the younger age of Mechanical Turk users but may also reflect higher education levels among early adopters of technology. Despite being more educated, Mechanical Turk workers report lower income. The shape of the distribution roughly matches the income distribution in the general U.S. population. However, it is noticeable that the income level of U.S. workers on Mechanical Turk is shifted towards lower income levels (U.S. Census, 2007). For example, while 45% of the U.S. Internet population earns below $60K/yr, the corresponding percentage across U.S.-based Mechanical Turk workers is 66.7%. (This finding is consistent with the earlier surveys that compared income levels on Mechanical Turk workers with income level of the general U.S. population (Ipeirotis, 2009). We should note that, despite the differences with the general population, on all of these demographic variables, Internet subject populations tend to be closer to the U.S. population as a whole than subjects recruited from traditional university subject pools.

Motivation. We asked respondents to report their motivations to participate in Mechanical Turk by selecting from a set of predefined options. Only 13.8% of the U.S.-based workers reported that Mechanical Turk was their primary source of income. However, 61.4% reported that earning additional money was an important driver of participation to the website. We should note though, that many workers also participate to Mechanical Turk for non-monetary reasons, such as entertainment (40.7%) and "killing time" (32.3%). In fact, 69.6% of the U.S.-based workers reported that they consider Mechanical Turk is a fruitful way to spend free time (e.g., instead of watching TV), a finding which is consistent with previous results (Chandler & Kapelner, 2010; Horton, Rand & Zeckhauser, 2010).

Most workers spend a day or less per week working on Mechanical Turk, and tend to complete 20–100 HITs during this time. This generates a relatively low income stream for Mechanical Turk work, which is often less than $20 per week. However, there are a few prolific workers that devote a significant amount of time and effort, completing thousands of HITs, and claim to generate an income of more than $1000/month (for a more detailed discussion see Ipeirotis, 2010). This reflects the substantial number of subjects in the U.S. who use Mechanical Turk as a supplementary source of income.

In sum, U.S. workers on Mechanical Turk are arguably closer to the U.S. population as a whole than subjects recruited from traditional university subject pools. Moreover, the increasing diversity of workers on Mechanical Turk makes it easy to conduct cross-cultural studies of decision making (e.g., Eriksson & Simpson, 2010). Note that Amazon allows participation in a given HIT to be restricted to workers from a specific country, allowing researchers to maintain a homogeneous population despite growing heterogeneity. In the following section we describe the features that configure Mechanical Turk as a sophisticated subject pool, and we address some potential concerns with its use such as validity and generalizability.

3 Conducting experimental research on Mechanical Turk

In this section we elaborate on how Mechanical Turk can serve as a sophisticated subject pool for running online experimentation. We point out some practical features that can make it easier to conduct certain kinds of research, and argue about how Mechanical Turk solves some typical concerns about Internet research.

3.1 Practical advantages of Mechanical Turk

Supportive infrastructure. Researchers who use Mechanical Turk benefit from the platform's services in various stages of the research process. Although the speed of recruitment depends on the HIT features (e.g. payment; Burhmester et al., in press), recruiting is generally fast. (It took us three weeks to collect 1000 subjects.) Moreover, because many workers accept to participate at the same time, Mechanical Turk makes it potentially simpler to run experiments that require interactions between subjects (e.g., game theory experimental designs or group decision-making).

When designing a HIT, researchers can either use Mechanical Turk's rudimentary in-house survey platform, or provide a link to an external site for workers to follow. When using external sites it is important to remember that eventually the experimenter will need to verify workers' claimed participation. One way to verify participation is to assign each worker an identification code (or have workers generate codes themselves) that can be used to match survey responses to payment claims. Requesters are paid by Mechanical Turk, possibly removing the burden of reporting individual payments (for tax purposes) from the hands of the experimenter. Therefore, in general the payment process is very smooth and accomplished with just a one-click procedure.

Subject anonymity. Workers are anonymous to the people who can view their responses. If subjects complete a HIT using external survey software, individual responses are not visible to requesters on Mechanical Turk, thus ensuring that subjects' responses cannot be linked to their identity by any single individual. Thus, Institutional Review Boards (IRBs, for research involving human subjects) are more likely to treat studies in Mechanical Turk as exempt from reviews, and this reduces concerns about how to safely store responses to sensitive questions.

Subject identifiability and prescreening. Mechanical Turk workers can be required to earn "qualifications" prior to completing a HIT. Qualifications are essentially prescreening questions that can be used to constrain who can see and complete particular HITs. Thus an experiment could be conducted on only women, or people who can correctly answer sports trivia questions, or people who are anxious about the economy, or whatever population the experimenter wishes to use. Qualifications can also be used to measure potential moderator variables, either by designing qualifications so that different responses lead different HITs to become visible (thus creating different groups that can then be compared) or by linking qualification responses to survey responses using the worker ID. This strategy allows a temporal separation between the collection of moderator variables and the collection of other variables, thereby reducing the likelihood that prescreening responses will contaminate subsequent responses.

Subject identifiability and longitudinal studies. Additionally, identifiability allows experimenters to continue collecting data from the same group of users over time. Worker IDs can be used to explicitly recontact former subjects or code can be written that restricts the availability of a HIT to a predetermined list of workers.

Cultural Diversity. HITs can be confined to only workers who live in specific countries, allowing for focused comparisons between subjects from two or more groups (Eriksson & Simpson, 2010). This can eliminate many of the barriers to conducting cross-cultural comparisons of

Table 1: Tradeoffs of different recruiting methods.

	Laboratory	Traditional web study	Web study with purpose built website	Mechanical Turk
Susceptibility to coverage error	High	Moderate	Moderate	Low
Heterogeneity of samples across labs	Moderate	High	High	Low
Non-response error	Low	High	High	Moderate
Subject Motivation	Moderate / High	Low	Low	Low
Risk of multiple responses by one person	None	Moderate	Moderate	Low
Risk of contaminated subject pool	Moderate	High	Moderate	Low
Risk of dishonest responses	Moderate	Low	Low	Low
Risk of experimenter effects	Low	None	None	None

basic psychological processes, namely, finding a subject pool in the country of interest or a collaborator who can collect the data. Furthermore, the content of each HIT posting can be uniquely tailored to the residents of that country.

This can allow subjects to see the survey in their first language (if desired), and decisions about money can be made using both the local currency and values that reflect the average wages and standard of living of that country. As Mechanical Turk is populated by an increasingly internationalized workforce, we foresee large scope for cross-culture comparisons in the future.

3.2 Potential threats to validity and generalizability.

As an unfamiliar method of recruiting subjects, researchers may be concerned about the validity and generalizability of results obtained from Mechanical Turk. There are two primary concerns about Mechanical Turk. First, there are concerns about whether Mechanical Turk workers are representative of the desired population as a whole (whatever that may be). Second, there were concerns about the overall quality of the data that respondents provide. We briefly review the reason why these issues are of concern and compare Mechanical Turk to other methods of data collection on these dimensions. Table 1 provides a comparative summary.

3.2.1 Representativeness of samples

Concerns about the representativeness of a sample include concerns about whether the people recruited and who choose to participate match the population of interest. Our demographic data suggests that Mechanical Turk workers are at least as representative of the U.S. population as traditional subject pools, with gender, race, age and education of Internet samples all matching the population more closely than college undergraduate samples and internet samples in general (see also Buhrmester et al., in press). More importantly, as we demonstrate in Section 4, non-response error seems to be less of a concern in Mechanical Turk samples than in Internet convenience samples recruited through other means.

3.2.2 Data quality

Given that Mechanical Turk workers are paid so little, one may wonder whether they take experiments seriously. Another concern is that the anonymity of the Internet may lead individual subjects to provide many separate responses to the same study. There is little evidence to suggest that data collected online is necessarily of poorer quality than data collected from subject pools (Krantz & Dalal, 2000; Gosling et al., 2004). Further, in practice, multiple responses are rare in web based experiments and are even less of a problem for Mechanical Turk because each worker ID must correspond to a unique credit card number (for a detailed discussion, see Horton et al., 2010).

One potential drawback of Mechanical Turk experiments (that actually applies to all web based experiments) is that unsupervised subjects tend to be less attentive than subjects in a lab with an experimenter (Oppenheimer et al., 2009). However, this problem is solvable; either through "catch trials" that identify subjects who failed to pay close attention, or through instructional manipulation checks that identify inattentive subjects and remind them to pay more attention (Oppenheimer et al., 2009).

3.2.3 Mechanical Turk can strengthen internal validity

Mechanical Turk workers can complete experiments without interacting with experimenters, possibly without even knowing that they are in an experiment. This avoids concerns of experimenter bias (Orne, 1962), subject crosstalk (Edlund et al., 2009) and reactance (for a detailed discussion of the validity of experiments conducted using online labor markets see Horton et al., 2010).

4 A comparative study

We have conducted numerous replications of traditional JDM findings on Mechanical Turk, suggesting that it is reliable (see http://experimentalturk.wordpress.com). We extend these findings by directly comparing Mechanical Turk data with data collected from other sources.

We recruited subjects from three different sources: Mechanical Turk, a traditional subject pool at a large Midwestern U.S. university, and visitors of online discussion boards. The study (carried out in April and May 2010) provides additional evidence on the consistency between Mechanical Turk workers and more traditional subjects, with respect to both actual behavior and attention provided to the experimental tasks.

4.1 The survey

Subjects completed three classic experimental tasks drawn from the heuristics and biases literature. The survey was completed using Qualtrics survey software. Questionnaires were identical across conditions with the exception that Mechanical Turk workers were asked to provide a code automatically generated by Qualtrics at the end of the experiment.

The *Asian disease problem* (Tversky & Kahneman, 1981) demonstrates framing effects. Subjects had to choose what action plan between a riskier and a safer one they preferred in order to contrast the outbreak of an unusual disease. In a between-subjects manipulation, the outcomes of the plans were either framed in positive terms (people saved), or in negative terms (people lost).

The *Linda problem* (Tversky & Kahneman, 1983) demonstrates the conjunction fallacy, that is the fact that people often fail to regard a combination of events as less probable than a single event in the combination. Respondents read a description of Linda and rated which of two alternative profiles was more likely to describe Linda, with one being more general than the other.

The *physician problem* (Baron & Hershey, 1988; Experiment 1, Cases 1 and 2) demonstrates the outcome bias, the fact that stated judgments of quality of a decision often depend on the valence of the outcome. Subjects rated on a seven-point scale (as used by Stanovich & West, 2008) the quality of a physician's decision to perform an operation on a patient. The operation was described as either a success or a failure in a between-subjects design.

After completing these three tasks, subjects completed the Subjective Numeracy Scale (SNS; Fagerlin et al. 2007). The SNS is an eight-item self-report measure of perceived ability to perform various mathematical tasks and preference for the use of numerical versus prose information. Because of its high correlation with the numeracy measure (Lipkus et al., 2001), the SNS provides a parsimonious measurement of an individual's quantitative abilities. Therefore, evidence of low quantitative score on the SNS may raise some concerns regarding the actual capacity of workers in Mechanical Turk to appreciate the magnitude of the wages/effort ratio in listed HITs.

Moreover, the SNS provided an ideal context for a catch trial that measured whether subjects were attending to the questions. Included with the SNS, subjects read a question that required them to give a precise and obvious answer ("*While watching the television, have you ever had a fatal heart attack?*"). This question employed a six-point scale anchored on "*Never*" and "*Often*" very similar to those in the SNS, thus representing an ideal test of whether subjects paid attention to the survey or not.

4.2 The samples

Amazon Mechanical Turk. We posted a task that required workers to complete an externally hosted survey in exchange for $0.10. The HIT was titled "Answer a short decision survey" and described as "Make some choices and judgments in this 5-minutes survey". The (overestimated) completion time was included in the HIT description in order to provide workers with a rough assessment of the reward/effort ratio. The actual ratio was $1.71/hour. The HIT was visible only to workers with an acceptance rate greater than 95% and who were residents in the U.S. One hundred thirty-one workers took part in the study.

Table 2: Subject pools characteristics.

Subject pool	% Females	Average age	Median age	Subjective numeracy (SD)	% Failed catch trials	% Survey completion
Mechanical Turk	75.0%	34.3	29	4.35 (1.00)	4.17%	91.6%
Midwestern university	68.8%	18.8	19	4.17 (0.81)	6.47%	98.6%
Internet boards	52.6%	30.6	26	4.25 (1.16)	5.26%	69.3%

Lab subject pool. One hundred and forty-one students from an introductory subject pool at a large Midwestern U.S. university completed this study.

Internet Discussion Boards. We posted a link to the survey to several online discussion boards that host online experiments in psychology. The survey has been available online for two weeks, and one hundred thirty-seven visitors took part in the study.

4.3 Results

Subjects' demographics. Subjects (N = 318, 66.0% female, M_{age} = 28.3) were recruited from Mechanical Turk, discussion boards around the Internet and an introductory subject pool at a Midwestern U.S. University. Subjects recruited from the Internet boards were comparable in terms of average age to subjects recruited from Mechanical Turk (34.3 and 31.8 respectively) and unsurprisingly, both were older than subjects recruited from the lab subject pool (18.8). Table 2 summarizes the demographics.

Non-response error. We looked at the number of people who accessed to the study but did not complete it entirely. As expected, almost everybody in the lab subject pool completed the study (98.6%). Subjects recruited from online discussion forums were significantly less likely to complete the survey than subjects on Mechanical Turk (66.7% and 91.6% respectively), $\chi^2(1,268)$ = 20.915, p < .001. This suggests that Mechanical Turk strongly diminishes the potential for non-response error in online research.

Attention. Subjects in the three subject pools did not differ in terms of attention provided to the survey. Subjects in Mechanical Turk had the lowest catch trial failing rate (defined as the proportion of subjects who did not select "*Never*" to the question "*While watching the television, have you ever had a fatal heart attack?*"), although the number of respondents who failed the catch trial is very low and not significantly different across subject pools, $\chi^2(2,301)$ = .187, p = 0.91. Subjects who failed the catch trial, or did not reach the page containing the catch trial, were removed from subsequent analyses.

Subjective numeracy. Subjects in the three subject pools did not differ significantly in the SNS score, $F(2,299)$ = 1.193, p = 0.30. As the SNS is closely associated with many measures of quantitative ability, this result suggests that workers in Mechanical Turk are not less able to handle quantitative information (e.g. payments for participation) than more traditional experimental subjects.

Experimental tasks. Table 3 summarizes the results obtained in the experimental tasks. The present tasks, along with their variations, are widely used in judgment and decision-making, and in particular they had already been posted repeatedly on Mechanical Turk (http://experimentalturk.wordpress.com; Horton et al., 2010). Therefore, for each task we excluded from the analysis subjects who declared they previously completed the task.

In the Asian disease problem, people were significantly more likely to choose the risky course of action when the problem was framed in terms of losses than when it was framed in terms of gains. Effect sizes are exactly the same across samples. Note that subjects on Mechanical Turk exhibited more risk aversion than subjects in the other subject pools, although this did not occur in previous tests of the same problem (http://experimentalturk.wordpress.com; Horton et al., 2010).

Respondents in all subject pools exhibited the conjunction fallacy. Large majorities regarded a combination of events ("Linda is a bank teller and is active in the feminist movement") as more probable than a single event in the combination (Linda is a bank teller"). We found slight differences across samples for this effect, $\chi^2(2,274$ = 4.606, p = 0.1, however this is consistent with the large variability of results in the conjunction fallacy literature (e.g., Charness, Karni, & Levin 2009).

Subjects in all the subject pools showed an outcome bias. In the physician problem, subjects judged the quality of the physician decision to be higher when it was followed by a success than when it was followed by a failure. The result is significant in all the subject pools, and the effect size in Mechanical Turk is the highest among the three samples.

Overall, these results confirm that Mechanical Turk is a reliable source of experimental data in judgment and decision-making. Results obtained in Mechanical Turk did not substantially differ from results obtained in a sub-

Table 3: Results on experimental tasks.

	Mechanical Turk	Midwestern university	Internet boards
Asian Disease			
% Risky Positive Frame	17.6%	28.1%	23.7%
% Risky Negative Frame	55.3%	67.7%	63.0%
χ^2	10.833	20.230	13.013
p	< 0.001	< 0.001	< 0.001
Effect size (w)	0.39	0.39	0.39
Linda problem			
% Conjunction Fallacy	72.2%	78.3%	64.4%
Physician problem			
Avg. Quality Success (SD)	5.93 (0.81)	5.63 (0.75)	5.73 (0.98)
Avg. Quality Failure (SD)	5.13 (1.24)	4.86 (1.29)	4.93 (1.41)
t	3.70	4.14	2.547
p	< 0.001	< 0.001	0.007
Effect size (d)	0.76	0.73	0.66

ject pool at a large Midwestern U.S. university. Moreover, response error was significantly lower in Mechanical Turk than in Internet discussion boards.

5 Concluding comments

Our theoretical discussion and empirical findings suggest that experimenters should consider Mechanical Turk as a viable alternative for data collection. Workers in Mechanical Turk exhibit the classic heuristics and biases and pay attention to directions at least as much as subjects from traditional sources. Furthermore, Mechanical Turk offers many practical advantages that reduce costs and make recruitment easier, while also reducing threats to internal validity. However, unlike traditional subject pools, which are reset every year when a new class arrives, Mechanical Turk membership evolves more organically and some workers may be potential experimental subjects for years. This means that experimenters will need to be more careful about how they manage relationships with subjects. We conclude the article highlighting two open issues that should be considered by experimenters in light of this difference.

Tracking subjects to ensure independent responses across experiments. Thanks to the workers' unique ID, researchers can identify workers who already took part to previous versions of an experiment, and exclude them accordingly. The easiest way to do this is to post a single HIT that redirects to a splash page. The url that the splash page directs workers to can be changed and all subjects will be unique. Researchers who have basic web programming skills can also specify workers that should not see the HIT, making it possible to post many HITs while avoiding subject pool contamination (for details on how to do this, see the most recent developer guide here: http: // developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=28)

However, there is no way to know whether a certain subject already took a similar version of the experiment posted by some other researcher. Given that many experiments are designed as slight variations of paradigmatic ones (e.g., Asian disease) it is probably also wise to ask subjects whether they already completed previous versions of the tasks.

Maintaining a positive reputation among workers. Experimenters should keep in mind that, although there is no direct way for workers to retaliate against poor employers, workers can refuse to complete tasks because the payment is clearly not adequate or because they previously had a bad experience with a requester. Because workers can share these bad experiences on blogs and other outlets (e.g., http://turkopticon.differenceengines.com), careless researchers can create problems for themselves (or others associated with them) if their HIT descriptions are confusing or misrepresent the time and effort required.

In principle requesters can offer workers wages that are disproportionately low, even considering what the norms are on Mechanical Turk, with little concern since data quality seems to be not affected by payments (Mason &

Watts, 2009). Workers are capable of sorting HITs by payment and reading the description before they choose to accept. However, researchers should be transparent about the wage they set and ensure that the time and the effort required to complete the task is provided in the HIT description. Not only is will this ensure that workers are able to make an informed decision about whether or not to complete a HIT, it will also reduce attrition (Crawford et al., 2001) and avoid potential reputational damage to the researcher.

A somewhat more complicated issue is deciding whether or not a HIT should be rejected. Here, the experimenter must balance community norms of Mechanical Turk, which require "bad" responses to be rejected, with the inherently subjective nature of experimental responses and IRB requirements to avoid penalizing subjects who withdraw from experiments. One possible solution is to include a non-subjective, non-experimental task before the survey that can be used to verify worker effort. Another solution is to require that subjects specifically click a link to opt out of the survey. In most cases, it may be best to give the workers the benefit of the doubt and pay them, but block them from participating in future experiments. When rejecting a HIT, it is always a good idea to provide concrete and justifiable reasons in writing to prevent misunderstandings.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (in press). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*, 569–579.

Buhrmester, M. D., Kwang, T., Gosling, S. D. (in press). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*.

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

Charness, G., Karni, E., & Levin, D. (2009). On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior*, *68*, 551–556

Crawford, S., Couper, M. P., Lamias, M. (2001). Web Surveys: Perceptions of Burden. *Social Science Computer Review, 19*, 146–162.

Edlund, J. E., Sagarin, B. J., Skowronski, J. J., Johnson, S. J., Kutter, J. (2009). Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk.*Personality and Social Psychology Bulletin*, *35*, 635–642.

Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, *5*, 159–163.

Fagerlin, A., Zikmund-Fisher, B., Ubel, P., Jankovic, A., Derry, H., & Smith, D. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making*, *27*, 672–680.

Gosling, S., Vazire, S., Srivastava, S., & John, O. (2004). Should We trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93–104.

Horton, J., & Chilton, L. (in press). The labor economics of paid crowdsourcing. *Proceedings of the 11th ACM Conference on Electronic Commerce*.

Horton, J., Rand, D., & Zeckhauser, R. (2010). The online laboratory: Conducting experiments in a real labor market. *NBER Working Paper* w15691.

Ipeirotis, P. (2009). Turker demographics vs. Internet demographics. http://behind-the-enemy-lines.blogspot.-com/2009/03/turker-demographics-vs-internet.html. Accessed August 18, 2010.

Ipeirotis, P. (2010). Demographics of Mechanical Turk. *CeDER-10–01 working paper*, New York University.

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). New York: Academic Press.

Lipkus, I., Samsa, G., & Rimer, B. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Mason, W., & Watts, D. (2009). Financial incentives and the "performance of crowds." *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation,* 77–85.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.

Rosenthal, R., & Rubin, D. B. (1979). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences*, *1*, 377–415.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI EA '10: Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pp. 2863–2872, New York, NY, USA. ACM.

Stanovich, K. E., & West. R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*, 672–695.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

U.S. Census Bureau (2007). Current Population Survey, 2007 Annual Social and Economic Supplement, Table HINC-06. http://pubdb3.census.gov/macro/032007/hhinc/new06_000.htm. Accessed August 19, 2010.