Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars

Kevin Crowston

Amazon Mechanical Turk: A research tool for organizations and information systems scholars

Kevin Crowston

Syracuse University School of Information Studies
348 Hinds Hall, Syracuse, NY 13210 USA
crowston@syr.edu

Abstract. Amazon Mechanical Turk (AMT), a system for crowdsourcing work, has been used in many academic fields to support research and could be similarly useful for information systems research. This paper briefly describes the functioning of the AMT system and presents a simple typology of research data collected using AMT. For each kind of data, it discusses potential threats to reliability and validity and possible ways to address those threats. The paper concludes with a brief discussion of possible applications of AMT to research on organizations and information systems.

Keywords: Amazon Mechanical Turk, crowd sourcing, research methods.

1 Introduction

The crowdsourcing system Amazon Mechanical Turk (AMT) was initially invented to support Amazon's business processes, but has since been used for research in disciplines such as natural language processing [e.g., 1, 2], machine learning [e.g., 3, 4] and human computer interaction [e.g., 5, 6]. There has been some social science use, e.g., in political science [e.g., 7] and psychology [e.g., 8]. However, AMT does not yet seem commonly used in information systems (IS) (one exception is Conley [9], who used AMT to code some of her data). To introduce AMT to IS researchers, I briefly explain how AMT works and present a simple typology of different applications of AMT to research. We then discuss concerns about reliability and validity of data generated by AMT for these different kinds of research. I conclude with suggestions for applying AMT to research on organizations and IS in particular.

AMT is a "marketplace for work that requires human intelligence" [10]. It provides a web-based system to dispatch tasks to a pool of human workers, known colloquially as Turkers. As such, it is an example of crowdsourcing, defined as outsourcing a function to a large by undefined group of people via an open call [11]. AMT is not the only crowdsourcing system, but it is well developed, commonly used and has the most information available to assess its suitability for research, hence our focus on it in this paper. In contrast to other systems for crowdsourcing, such as InnoCentive, the tasks on AMT are typically small (i.e., a few minutes to perform rather than days or weeks), and payments are low (on the order of a few cents).

As a background to our discussion of research uses of AMT, we first briefly walk through the steps involved in using the system. Interested readers may wish to consult Mason & Suri's [12] paper on using AMT for behavioural research.

The unit of work done on AMT is called a human intelligence task (HIT). A HIT may be carried out entirely on Amazon's system, e.g., for a survey, or may refer the Turker to another website for more complicated tasks, e.g., an on-line experiment. The most common HITs on AMT are tasks such as transcription, content generation or classification of images for companies, mirroring the original purpose for Amazon [13]. Research tasks are a small part of the total volume of HITs: they add novelty for Turkers but are not an important concern for Amazon.

Most HITs offer very small payments. Ipeirotis [13] reports that "25 percent of the HITs created on Mechanical Turk have a price tag of just US$0.01, 70 percent have a reward of $0.05 or less, and 90 percent pay less than $0.10" (p. 19). The payment offered is expected to roughly reflect the difficulty of the task; if desired, a bonus can be paid above the base amount. Horton & Chilton [14] estimated the reservation hourly wage ("the minimum wage a worker is willing to accept… for performing some task") to be about US$1.40/hour (p. 2). The actual pay offered seems to be somewhat higher: Ipeirotis [13] reports an average pay of US$4.80/hour for tasks (high enough that Turkers in low-wage countries can earn a living from AMT). Higher pay may motivate Turkers to work on tasks more quickly, though Buhrmester et al. [8] found that even tasks that paid lowest amount (US$0.01) did eventually attract some Turkers, apparently without reducing data quality. Contrariwise, a too high payment (e.g., more than US$1) is viewed by Turkers as signalling a bogus task [15]. Amazon charges 10% on top of the amount paid to the Turker.

The creator of a HIT determines the number of Turkers wanted and how many times an individual Turker can respond. For example, a survey should allow a Turker to respond only once, while a classification task might require 3 different responses for each item but allow the same Turker to classify multiple items. AMT enables some limited screening of Turkers. HITs can be restricted to Turkers in particular countries, which may be helpful for cross-cultural studies. A HIT may require that Turkers be prequalified, e.g., by requiring acceptable answers on prescreening questions before taking a survey or doing an experiment. Turkers with less than a given percentage of satisfactorily completed tasks can be blocked. Turkers who have completed a previous HIT can be invited to a new one, e.g., for a panel survey.

Once a HIT is released, it goes into long list of available HITs: on the order of 100,000 HITs may be available at a time [15]. Amazon presents a browsing interface with hundreds of pages of HITs, as well as a simple search interface. Chilton *et al.* [15] found that the first page of HITs has the highest "click through" rate and that most Turkers use the "recently posted" and "largest number of HITs" sort orders (the later to be able to repeat a HIT). Research work is likely to be prominent only with the first sort order. HITs are typically completed very quickly, within a day or two. However, if a HIT is not completed within that time frame, it may be delayed for a long time since it will disappear from first few pages of the recent post list [13].

Results provided by Amazon include answers to the questions posed, an ID that allows tracking results from same Turker and time spent on the task. After reviewing

results, a poster can choose to not pay or to block a particular Turker from future tasks, e.g., if the work done is of low quality. The possibility that they might not get paid leads to questions among Turkers about the honesty of new posters. Turkers talk amongst themselves on a variety of web forums, so a reputation for not paying (or contrary) will spread [16].

The AMT system offers many potential benefits for research. First, the cost of recruiting subjects is much lower than alternatives [7], e.g., US$0.50 per subject rather than a $10 gift card. Second, Amazon handles all payments and Turkers are anonymous to researchers. Third, the large pool makes it easy to recruit a diverse set of subjects and to get multiple subjects to participate at same time, e.g., for group experiments [12]. Finally, work is typically done within hours or days [13].

However, AMT has significant limitations as a research environment. Since the work is being done remotely, the researcher has no control over the physical environment (e.g., what setting or what kind of computer and monitor) or even the web environment (e.g., which browser). There are only basic features for selecting and filtering participants. It is difficult to know how well Turkers understand a task or how hard the Turker will concentrate on it. There are only a few very limited ways to follow up with a Turker after a task (e.g. sending a bonus or a payment rejection with a message); the only possibility for debriefing is to include questions in the HIT [17]. These limitations mean that certain kinds of studies will not be not feasible (e.g., these limits seem to pose significant challenges for interpretivist research). And of course, the novelty of the system creates concerns about the reliability and validity of the data, which we discuss in the following section.

2 Research Data from Amazon Mechanical Turk

Considering the subject of the data collection, we suggest that there are three general ways to use AMT to collect research data (as shown in Figure 1):

1) to collect data about Turkers,
2) to use Turkers to collect data about some research stimulus, or
3) to collect data about the Turkers' reactions to some stimulus.

Published papers using AMT generally examine only a single mode of data collection, without being explicit about how the characteristics of the task affect the

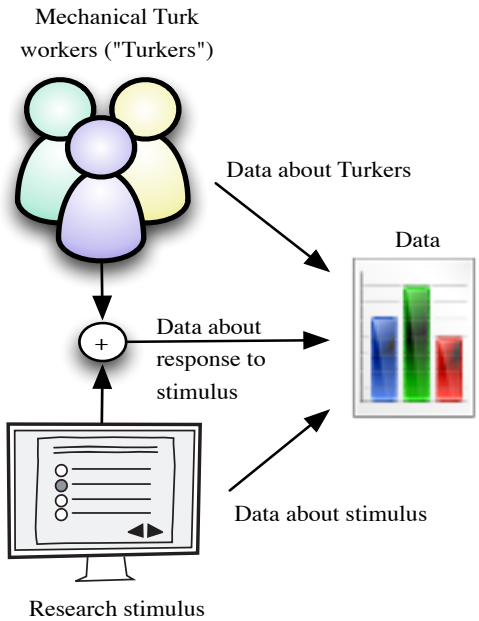

Fig. 1. Three modes of data collection using AMT.

outcomes. However, it is useful to distinguish these modes, as the nature of the different data collected raise unique reliability and validity concerns with different possible remedies. For each mode, we consider reliability (i.e., random error in the data), internal validity (i.e., bias in the data) and external validity (i.e., the possibility to generalize from the data).

2.1 Data about Turkers

The first possibility is to use AMT where the data to be collected are about the Turkers themselves, e.g., using Turkers as subjects for a survey or psychological test [e.g., 5, 16]. In this case, an individual Turker will respond only once to a HIT. However, the same Turker might take part in different HITs, e.g., for repeated runs of an experiment. If multiple responses are undesirable, they can be removed from analysis or prevented in the implementation of the HIT.

Reliability of the data collected can be addressed as in other survey designs, e.g., by including multiple items for each construct (though this approach is more appropriate for positivist than for interpretivist research approaches). A more difficult concern is the internal validity of the data, i.e., if the respondents reply truthfully rather than simply making up data. Mason & Suri [12] reported that only a small percentage of respondents to their survey seem to be falsifying demographic data, and report similar results from other researchers. However, others have found that a few Turkers—referred to as spammers [12]—will try to do a HIT as quickly as possible without attention to the instructions or will even game the system, e.g., by quickly submitting many erroneous answers [4, 6]. A one-time survey offers a limited opportunity for spam, but as surveys can be quick to fill out (especially if one skips reading the questions), the possibility needs to be guarded against.

A simple spam detection approach is to look at the time taken to perform the task (e.g., a survey answered in a minute) or at the pattern of responses (e.g., all questions answered with the same option). A survey can include a check question to determine if respondents are actually reading the questions. For example, Paolacci et al. [16] included the question, "While watching the television, have you ever had a fatal heart attack?", using a six-point scale ("Never" to "Often") identical to that used for other questions. Any survey that was not answered "Never" was discarded. They observed about the same failure rate on this check question (on the order of 5%) in AMT as with other subject pools (students and Internet pools).

A second concern about using AMT to recruit subjects is the question of external validity. Clearly, using AMT is not a substitute for surveying members of an organization, but Turkers might be taken as representative of the general population of Internet users more specifically. There has been research on the demographics of the population of Turkers that addresses this question, revealing that Turkers in fact differ somewhat from the reported averages for Internet users in general [18]. Overall, Turkers were younger than average Internet users. The self-reported education was higher than average, but income lower. Most were single and without children. Furthermore, there are differences within the pool of Turkers, with resulting variability in other capabilities, e.g., the level of English ability [3]. Ipeirotis [18]

reported finding Turkers from 66 countries, but 46% of Turkers were US-based and 34% Indian (the two currencies in which Amazon offers payment). US Turkers were about 2/3 female, while Indian Turkers were about 70% male [18]. Money was an important factor for all Turkers, but US Turkers generally viewed AMT as a secondary source of income, while for many Indians it was a primary source of income. The importance of a monetary incentive for participation could be a contaminating factor for studies of motivation or related phenomena, though this problem also affects studies done with other paid subject pools.

Despite these differences, the Turker population may not be appreciably less representative of the Internet or general population than other commonly-used subject pools, such as college students or subjects recruited on the Internet. Buhrmester et al. [8] noted that the AMT sample is likely to be more diverse than the others. Berinsky et al. [7] found that a sample of AMT respondents was at least as representative, if not better than convenience samples or student samples and was most similar to a probability Internet survey. The sample was different from a face-to-face probability sample of US participants, but was described as not "wildly distorted" (p. 12).

Finally, in this case, the Turkers are human subjects for the research, so the rules and ethical principles that govern human subjects research apply, e.g., the requirement for obtaining informed consent before starting the research task and balancing risks and benefits of the research. It seems likely though that the risks of participating in an AMT HIT would be minimal. Indeed, use of AMT might reduce some risks, e.g., the anonymity of the system would reduce the risk of inadvertent disclosure of private data, which may be beneficial for human subjects review [16]. An issue to consider is if the use of AMT shifts the burden of the research to a vulnerable group, i.e., to those willing to work for the low pay offered. As always, careful consideration of the risks and benefits of the research is needed.

2.2 Data about a Research Stimulus

A second possibility for AMT research is that the researcher is studying some collection of objects that need humans to provide data about them. We refer these objects generically as research stimuli; an individual Turker may work on many stimuli. In this case, we assume that the data of interest are inherent in the stimuli (i.e., different individuals are not expected to have different interpretations) and require observation with minimal interpretation. (These assumptions are more characteristic of a positivist research approach.)

This case describes many uses of AMT for research. For example, Kaisser & Lowe [2] had Turkers read documents known to contain the answer to a question; the Turkers identified the specific sentence in the document that included the answer. Wang, Kraut, & Levine [19] had Turkers code discussion forum messages for whether they offered information or emotional support. Sorokin & Forsyth [4] had Turkers annotate images to locate people.

A first issue with this mode of data collection is reliability (i.e., random errors in the data). Researchers have identified a variety of approaches to resolve this issue. First, tasks must be carefully designed, since Turkers have only the training offered in

the HIT. For example, Sorokin & Forsyth [4] explored four alternative task designs to identifying people in pictures to determine which could be done most reliably. A second strategy is to require a qualification task to ensure that subjects understand the task and can (and will) do it [5]. Rashtchian et al. [3] reported that pre-screening Turkers led to the highest improvement in quality on an image annotation task.

A third common strategy is to use replication, i.e., instead of doing content analysis with a few trained analysts, use more untrained workers, pooling data to obtain a consensus for each stimulus. Snow et al. [1] found that combining judgments from about five Turkers on factors such as "emotions expressed, the relative timing of events referred to in the text, word similarity, word sense disambiguation, and linguistic entailment or implication" [19] gave results similar to experts.

Finally, Turkers can be used to validate work done by other Turkers, i.e., create a HIT that presents the prior task and its responses and asks Turkers to judge whether the response is appropriate for the task. However, researchers suggest that it is more effective to use other strategies to ensure higher initial reliability rather than trying to filter out bad work after the fact [3].

The second issue is internal validity, i.e., does the data that non-trained Turkers can extract really represent what the researcher wants to study? To guard against inadvertent bias requires careful design of the stimulus and instructions (as in any research setting). Sprouse [17] obtained essentially identical results from an AMT experiment and a laboratory experiment on a linguistic judgement task, suggesting that AMT experiments can provide valid data.

Spam is a more significant issue in this mode of data collection, since a Turker can work on multiple tasks. Ipeirotis, Provost, & Wang [20] estimated that 30% of the responses to a task they posted were provided by spammers; the spammers were a small number of the total, but posted many bogus responses. Therefore, the HIT must be designed to deter and enable detection of spammers. As Kittur et al. [6] put it, the task should be designed "such that completing it accurately and in good faith requires as much or less effort than non-obvious random or malicious completion." (p. 456).

The strategies used for the previous mode can be used in this case as well, i.e., checking the timing and pattern of responses or asking a question that demonstrates that workers are paying attention to the task. Another simple approach to spam detection is to include a few stimuli with known correct answers ("gold standard" data); responses to these stimuli can be used to check the quality of a Turker's work. Responses from multiple Turkers can be compared to detect Turkers who are outliers. For example, Sprouse [17] plotted the distributions of data from different Turkers and rejected those that were significantly different from the others (about 14.2% of the total responses). Ipeirotis, Provost, & Wang [20] developed a more sophisticated model for detecting quality of Turkers on a classification task that requires 5 labels per object and 20-30 objects per Turker.

Finally, because the data are not about the Turkers themselves, the issue of representativeness and external validity of Turkers as a sample does not arise (though there may be issues concerning the representativeness of the sample of stimuli). A further consequence is that the ethical concerns regarding the use of human subjects in research do not apply. Instead, the Turkers can be seen as out-sourced employees,

raising a different set of concerns about the fairness of such employment [e.g., 21]. One concern might be about the quality of the job offered, though as noted above, the typical wage offered is low only by the standards of developed countries; it is considerably more than the average in most developing countries.

2.3 Data about Reactions to a Research Stimulus

The final possibility is that the data of interest are not about the Turkers themselves nor explicit in a stimulus but rather come from interaction of people with a stimulus. It might be argued that data in mode 1 also come from an interaction with a stimulus such as a survey or test, but the difference here is that we are interested in data about the stimulus, not just the Turker. For example, a common research use of AMT is to recruit users for tests of IT systems in order to get usage data and user feedback [e.g., 6]. The goal of the study may be to compare different stimuli or to determine how members of different groups react to the same stimulus (e.g., which kinds of users find a particular system harder to use). In these cases, subjects' responses to the stimulus are expected to be different, rather than simply reflecting an underlying truth inherent in the stimulus as in mode 2. As a result, this mode of data collection presents the most challenging issues for both validity and reliability.

Validating subjective data for reliability is inherently difficult. Some of the techniques from the other modes may carry over. As in mode 1, it may be possible to use multiple items per construct to assess reliability. As in mode 2, careful task design and prequalification of Turkers will be useful. However, since many different answers could plausibly be correct [6], it is not possible to use "gold standard" data, to spot check results or to use replication to arrive at a consensus. These limitations would seem to limit the usefulness of AMT for interpretivist research in particular.

Spam continues to be a possible threat to internal validity. One approach is to include a few questions that can be used to check that the work required for the task is actually being performed, even if the work itself can not be checked. For example, Kittur et al. [6] had Turkers evaluating the quality of a Wikipedia page also report on "how many references, images, and sections the article had. In addition, users were required to provide 4-6 keywords." The answers to the first three questions were used to verify that the Turker had actually viewed the page; the answer to the last question required the Turker to carefully read the page, as required to rate quality.

This mode of data collection poses additional threats to internal validity. Berinsky et al. [7] suggested that because of concern about getting paid, conscientious Turkers may follow instructions closely, resulting in higher risk of researcher demand. For example, in evaluating a system, participants may provide positive or enthusiastic responses under the assumption that this will improve their chances of getting paid. It is also possible for an experiment that Turkers will discuss the experimental conditions in message boards or through other means [12]. However, if the experiment concludes quickly, this may not be a practical problem.

The specific demographics of Turkers raise concerns about the external validity of studies, as discussed above. However, Paolacci et al. [16] repeated three well-known psychological tests with Turkers and obtained results comparable to prior results,

again suggesting that AMT results can generalize. Finally, as in the first case, since data will be collected about the Turkers, they will likely be considered human subjects in the research and so the concerns about the use and protection of human subjects apply.

To summarize, the specific recommendations made above to address concerns of reliability and validity in the three cases are presented in Table 1.

Table 1. Summary of recommendations to address reliability and validity of data from different modes of data collection.

Research concern	**Mode 1: Data about Turkers**	**Mode 2: Data about research stimulus**	**Mode 3: Data about interaction**
Reliability (i.e., errors in responses)	Use multiple indicators per construct	Prequalify Turkers Replicate work Use AMT to validate responses	Use multiple indicators per construct Prequalify Turkers
Internal validity (i.e., biased responses)	Prevent or remove duplicate responses Consider effects of monetary compensation on research questions		Same as mode 1 Design task to minimize demand Minimize time to reduce discussion of experiment
Spam	Examine time taken to perform task Examine pattern of responses Include check questions	Same as mode 1 Include gold standard data Compare responses to detect outliers	Same as mode 1 Include objective-answer questions that demonstrate task performance
External validity (i.e., generalizability)	Not perfectly representative of Internet users, but not worse than alternatives	N/A	Same as mode 1

3 Case Study

To illustrate the application of AMT to an Information Systems research project, we present an example drawn from our own research. In this presentation, we present only how we used AMT to conduct the research; the details of the research questions, theories and the results of the study are reported elsewhere [22].

We have been conducting a design science research project to design and build a new citizen science system. In citizen science projects, members of the public are recruited to contribute to scientific investigations [23, 24]. Our project addresses a challenging problem in the life sciences: the taxonomic classification of plant, animal,

and insect species from photographs. A photograph of a specimen, tagged with the date and location where it was taken, can provide valuable scientific data (e.g., on how urban sprawl impacts local ecosystems or evidence of local, regional, or global climactic shifts). To be useful though, it is necessary to know what the picture is of, expressed in scientific terms, i.e., the scientific name of the species depicted. *Citizen Sort* was developed to let members of the public view collections of pictures maintained by researchers and annotate them with data about the specimens they depict, with the goal of classifying the picture as a particular species. To motivate participation, we drew on the idea of purposeful gaming, developing *Happy Match*, a sorting and matching game that awards points and high scores for classification. To be successful, the game needs to motivate users to both play and to create quality data about the photographs.

We used AMT to conduct an initial evaluation of the game. Our study falls into the third category above: we are interested in the reactions of Turkers to the system as a kind of research stimulus (design science system evaluations would generally follow this pattern). We note that the AMT subject pool is not really appropriate to test theories about motivation, as offering payment makes it difficult to assess the effects of other motivations. However, in this preliminary evaluation our main interest was on the question of data quality (could untrained users successfully classify photographs), as well as the general usability/playability of the system. The possibility of rapid results offered by AMT seemed a good tradeoff for coverage of all research questions for this stage of the project. As well, AMT users seemed to be representative of our target population of active Internet users.

In setting up the HIT, we offered to pay participants US$0.50 for playing the game and completing a survey. To motivate good performance on the game, we offered a bonus of US$0.50 for getting a high score on any round in the game. We linked performance on the game to the survey results using a unique identifier, though a few players did not copy the identifier correctly, making their data unavailable for analysis. We offered to pay up to 100 users in each round of the study and ran two rounds in total, for a planned total of 200 participants. Because of the way AMT works, more than 100 people started each round. However, not all who started completed the task and of those who did, not all completed the survey that was necessary to be paid.

Those who accepted the AMT task were asked to accept an informed consent statement, to play *Happy Match* at least once and to then fill out the survey. The *Happy Match* system collected the number of games each player played and their score on each game. From the scores, we computed both the average score and high score. Finally, the system recorded each classifications performed by the users. For this initial evaluation, we only used photographs for which we had a professionally applied classification, enabling us to check the agreement of every user classification decision with the known data. From these data we computed each player's overall accuracy (the fraction of their classification that agreed with the expert), which we used to explore factors affecting data quality. After playing, users filled out a 28-item survey administered through AMT; these data were used to identify which users were more or less accurate as well as to explore motivational factors.

The results of our trial show both the strengths and weaknesses of the use of AMT. On the positive side, we were able to recruit a large number of users in a very short time and at low cost. For each round, we had the desired 100 responses within a day at a total cost of less than $100.00 per round. The subject pool was also much more diverse than a student pool would have been (e.g., ages ranged from 18 to 65). On the negative side, a few of the participants were apparently spammers, making little effort to play the game or to answer the survey questions; their data had to be filtered from the results. (We still paid them, as US human subjects rules require that we provide the offered compensation to any subject who starts the research.)

4 Conclusion

Prior experience with AMT suggests that with careful task design, AMT offers an interesting new capability to recruit research subjects or labour for a research project, providing useful research data. The typology presented above suggests relevant approaches. For studies of information systems and organizations, the first mode of data collection noted above is likely to be of limited use, as Turkers are likely too general a population for organizationally-focused research. Still, they may be a reasonable sample for studies of Internet use in general. Researchers could use this subject pool to examine attitudes or beliefs about technologies or specific systems. For example, a survey could be directed to users of eCommerce sites, such as Amazon, to examine attitudes or beliefs about the site's features or security.

AMT can provide a pool of workers to analyze research data, the second mode of use. It may be possible to crowdsource certain kinds of qualitative data analysis (e.g., content analysis), using the large number of Turkers to offset their minimal training. For example, researchers might use AMT to code email messages for evidence of particular kinds of group processes to explore how different kinds of participation is related to group effectiveness. A concern specific to organizational research is how to protect confidential data when its analysis is crowdsourced. However, many companies use AMT for their data, suggesting that this problem can be addressed.

Finally, studies in the third mode are likely to be particularly interesting for design science researchers, who might use AMT to recruit pilot study participants for system evaluations. As shown in the case study, the author has had some success using AMT in this way for a quick evaluation of a design science prototype. AMT could also be used for experiments by randomly assigning participants to different conditions. For example, a researcher could test the merits of an innovative interface by comparing the performance of Turkers on a task using a new and current system interface.

Acknowledgements

This research was partially supported by US NSF Grant 09–68470. The paper has benefited from helpful comments from Nathan Prestopnik and three anonymous reviewers.

References

1. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—But is it good?: Evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254–263. Association for Computational Linguistics, (2008).
2. Kaisser, M., Lowe, J.: Creating a research collection of question answer sentence pairs with Amazon's Mechanical Turk. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08). (2008)
3. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139–147. Association for Computational Linguistics, (2010).
4. Sorokin, A., Forsyth, D.: Utility data annotation with Amazon Mechanical Turk. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8. (2008), doi: 10.1109/CVPRW.2008.4562953
5. Heer, J., Bostock, M.: Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI '10), pp. 203–212. ACM, (2010), doi: 10.1145/1753326.1753357
6. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with Mechanical Turk. In: Proceedings of the ACM Conference on Human-factors in Computing Systems, pp. 453–456. ACM New York, NY, USA, (2008)
7. Berinsky, A.J., Huber, G.A., Lenz., G.S.: Using Mechanical Turk as a subject recruitment tool for experimental research. Political Analysis (2011),
8. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's Mechanical Turk. Perspectives on Psychological Science 6, 3–5 (2011), doi: 10.1177/1745691610393980,
9. Conley, C.A.: Design for quality: The case of Open Source Software development. PhD dissertation. New York University, New York, NY (2008)
10. http://www.amazon.com/gp/help/customer/display.html?nodeId=16465291
11. Brabham, D.C.: Crowdsourcing as a model for problem solving: An introduction and cases. Convergence 14, 75 (2008),
12. Mason, W., Suri, S.: Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods 44, 1–23 (2012),
13. Ipeirotis, P.G.: Analyzing the Amazon Mechanical Turk marketplace. XRDS 17, 16–21 (2010), doi: 10.1145/1869086.1869094,
14. Horton, J.J., Chilton, L.B.: The labor economics of paid crowdsourcing. In: Proceedings of the 11th ACM Conference on Electronic Commerce, pp. 209–218 (2010), arXiv:1001.0627v1
15. Chilton, L.B., Horton, J.J., Miller, R.C., Azenkot, S.: Task search in a human computation market. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 1–9. ACM, (2010), doi: 10.1145/1837885.1837889
16. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on Amazon Mechanical Turk. Judgment and Decision Making 5, 411–419 (2010),
17. Sprouse, J.: A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. Behavior Research Methods 43, 155–167 (2011), doi: 10.3758/s13428-010-0039-7,
18. Ipeirotis, P.G.: Demographics of Mechanical Turk. Working Paper CEDER-10-01, New York University (2010), http://ssrn.com/abstract=1585030
19. Wang, Y.-C., Kraut, R., Levine, J.M.: To stay or leave? The relationship of emotional and informational support to commitment in online health support groups. In: Proceedings of

the ACM Conference on Computer Supported Cooperative Work, pp. 833–842. ACM, (2012), doi: 10.1145/2145204.2145329
20. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on Amazon Mechanical Turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 64–67. ACM, (2010), doi: 10.1145/1837885.1837906
21. De George, R.: Information technology, globalization and ethics. Ethics and Information Technology 8, 29--40 (2006),
22. Crowston, K., Prestopnik, N.R.: Motivation and data quality in a citizen science game: A design science evaluation. In: Proceedings of Hawai'i International Conference on System Science. (2013)
23. Cohn, J.P.: Citizen science: Can volunteers do real research? BioScience 58, 192-107 (2008),
24. Wiggins, A., Crowston, K.: From conservation to crowdsourcing: A typology of citizen science. In: Proceedings of 44th Hawaii International Conference on System Sciences, pp. 1-10. (2011), doi:10.1109/HICSS.2011.207