Exploring Crowd Consistency in a Mechanical Turk Survey

Peng Sun
Department of Computer Science
North Carolina State University
psun2@ncsu.edu

Kathryn T. Stolee
Department of Computer Science
North Carolina State University
ktstolee@ncsu.edu

ABSTRACT

Crowdsourcing can provide a platform for evaluating software engineering research. In this paper, we aim to explore characteristics of the worker population on Amazon's Mechanical Turk, a popular microtask crowdsourcing environment, and measure the percentage of workers who are potentially qualified to perform software- or computer science-related tasks. Through a baseline survey and two replications, we measure workers' answer consistency as well as the consistency of sample characteristics. In the end, we deployed 1,200 total surveys that were completed by 1,064 unique workers. Our results show that 24% of the study participants have a computer science or IT background and most people are payment driven when choosing tasks. The sample characteristics can vary significantly, even on large samples with 300 participants. Additionally, we often observed inconsistency in workers' answers for those who completed two surveys; approximately 30% answered at least one question inconsistently between the two survey submissions. This implies a need for replication and quality controls in crowdsourced experiments.

1. INTRODUCTION

Crowdsourcing has become a common approach for conducting empirical studies in software engineering [5,6,18–20], and further it is being used to explore how to accomplish various software engineering activities, such as program synthesis [1], program verification [16], and testing [2,12]. The quality of the crowd's work, then, has a strong impact on research results and potentially also on code quality. However, response quality can be difficult to measure and control in a crowdsourcing environment, particularly one in which a large number of participants are completing many small tasks, as is the case with microtask crowdsourcing.

In this paper, we explore the characteristics of workers on Amazon's Mechanical Turk (MTurk), a microtask crowdsourcing platform. This tool provides an environment for deploying large number of tasks for completion by workers for pay. As many software engineering researchers are turning to crowdsourcing to complete tasks and evaluations (e.g., [5, 6, 10, 19, 20]), our goal was to measure what percentage of the MTurk workforce is potentially qualified to complete software engineering related tasks. By deploying the same, 8-question survey three times (i.e., one baseline survey and two replications) and allowing workers to participate twice, we were able to also measure within-participant consistency on simple questions related to age, gender, and education. Our results have implications for researchers who design and run studies on MTurk.

Our contributions are:

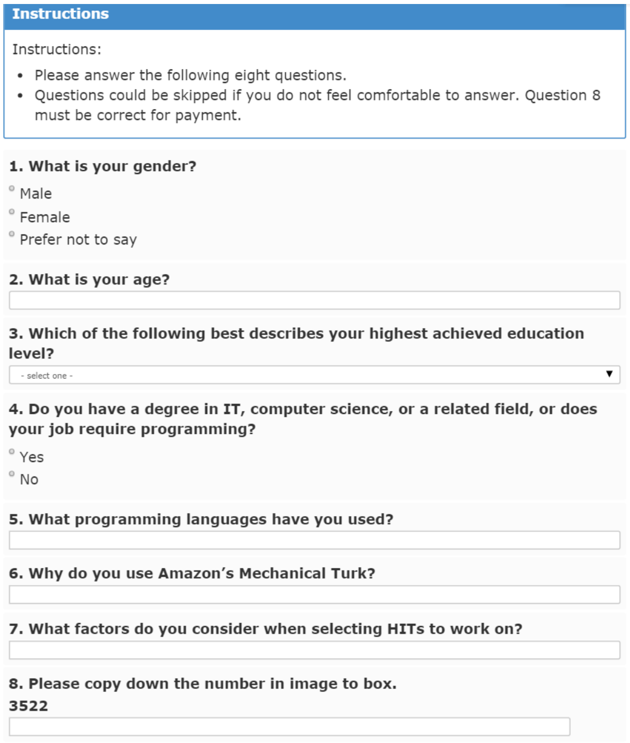
- Analysis of 1,200 survey responses from 1,064 unique MTurk participants regarding gender, age, computer science experience, and motivations for participating in MTurk,
- Comparison of characteristics of large (n = 300), independent samples from the MTurk worker population, showing some significant differences in computer science experience and gender, and
- Within-participant consistency analysis for 136 MTurk participants, illustrating inconsistency in responses for 30% of the workers.

2. BACKGROUND AND RELATED WORK

The MTurk crowdsourcing platform provides a low barrier to entry for people (requesters) to obtain access to a large pool of people (workers) to perform human intelligence tasks (HITs). Requester are those who create tasks and workers are those who complete tasks. The complexity of the HITs varies from labeling an image to completing domain-specific tasks such as evaluating source code [20]. Workers get paid if their assignment gets approved by the requester.

Many researchers have surveyed the characteristics of the MTurk worker pool, asking questions related to gender, age, education level, and motivation [4, 11, 13–15]. These efforts complement our own in seeking to understand more about who the MTurk workers are and why they participate. Paolacci, et al. [13] show that among U.S.-based workers, there are more females (64.85%) than males (35.15%) and the average age is 36. Ross, et al. [15] report similar demographic information in four studies from Nov 2008 to Nov 2009. Among U.S. workers, between 63% and 72% are female and between 28% and 37% are male. Average age varied from 33.2 to 35.4. Among workers from India, the ratios are reversed with 25%-39% female and 61%-75% male. Average

CSI-SE'16, May 16 2016, Austin, TX, USA
© 2016 ACM. ISBN 978-1-4503-4158-5/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2897659.2897662



Instructions

Instructions:

- Please answer the following eight questions.
- Questions could be skipped if you do not feel comfortable to answer. Question 8 must be correct for payment.

1. What is your gender?

○ Male
○ Female
○ Prefer not to say

2. What is your age?

3. Which of the following best describes your highest achieved education level?

- select one - ▼

4. Do you have a degree in IT, computer science, or a related field, or does your job require programming?

○ Yes
○ No

5. What programming languages have you used?

6. Why do you use Amazon's Mechanical Turk?

7. What factors do you consider when selecting HITs to work on?

8. Please copy down the number in image to box.
3522

Figure 1: Survey questions and instructions

age varied from 26.4 to 28.8. Our survey showed a more even split of males and females, with the exception of Study 2, though unlike prior work, we did not measure or control for geographic location (Table 2).

Other research in understanding crowds focuses on economic factors and incentives. For incentives, Yin, et al. [21] show that Mturk workers are most motivated by financial incentives and their own potential to perform well on the task. Paolacci, et al. [13] report that 75.2% of U.S.-based workers are motivated by money. Our study echoes these findings, showing that between 83% and 87% of workers are motivated by money. We note, also, that the relationship between money and performance has been studied. Finnerty, et al. [4] show that dynamic rewards (rewards paid based on performance including accuracy and time taken to complete the task) lead to better performance. Mason and Watts report no relationship between increased financial incentives and work quality on MTurk [11].

One of the most important threats to the integrity of crowdsourced activities, particularly for microtask crowdsourcing, is random clickers who complete tasks mindlessly. Researchers have tried to identify whether participants answer questions honestly in crowdsourcing study with different ways [3,8,9]. Downs, et al. use two questions of varying difficulties in a survey, and show that 39% of participants did not answer conscientiously [3]. This percentage is a little higher than our results (30%), which may be due in part to differences in study designs (Section 3).

3. STUDY

This study was designed as a survey to be deployed on MTurk. The survey is intended to help us understand what percentage of MTurk participants could be potentially qualified to complete tasks related to computer science and programming. Shown in Figure 1, it collects basic demographic information, education information, and motivations. There are four multiple-choice questions about gender, age[1], education level, and background in CS, IT, or a related field. There are three open-ended questions about programming languages, motivations, and task selection factors. The final question was used for quality control to ensure participants were paying some attention to question content. Accuracy on Question 8 was required for payment.

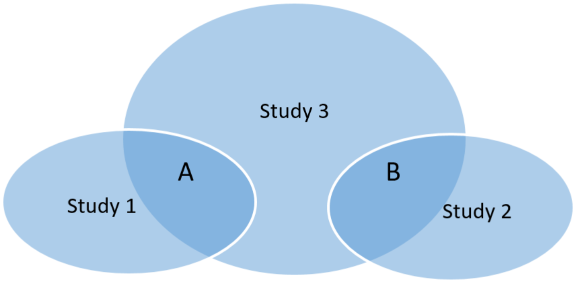


Figure 2: Study Design with Worker Overlap

This survey was deployed three times in total, one baseline study and two replications. As the general intention of replications is to verify results, we characterize the functions of the replications using the classification presented by Gómez, et al. [7]. The baseline survey (Study 1) was deployed to 300 MTurk participants on August 2, 2015. To understand population limits and the extent to which the results (sample characteristics) hold for other subjects, we deployed an exact survey replication to an independent sample of 300 MTurk participants two weeks later on August 16 2015 (Study 2). This allows us to compare characteristics across samples. To control for sampling error, we increased the sample size to 600 participants and deployed a second replication (Study 3) on August 23, 2015. Study 3 was open to all MTurk workers, regardless of past participation in Study 1 or Study 2. By allowing participants to complete the survey a second time, this provided an opportunity to look for within-participant consistency in survey responses. Figure 2 shows a visualization of the three studies, labeling the potential overlap in participants as *A* and *B*.

Our survey and study were designed to answer the following research questions:

RQ1: What percentage of MTurk participants have a computer science or IT background?.

This was evaluated using Questions 4 and 5 in the survey. The goal was to gain insight into what percentage of the MTurk population is potentially qualified to perform software engineering related tasks, and in what languages.

RQ2: What are the participants' motivations for using crowdsourcing platform Amazon's Mechanical Turk?.

[1] Note that while question 2 on age appears to be free-text, the contents of the text box required a numeric response.

This was evaluated using Questions 6 and 7 in the survey. The goal was to understand ways to incentivize workers when recruiting participants on MTurk.

RQ3: How consistent are the characteristics of samples of participants on MTurk?.

This research question was evaluated by comparing the survey responses on the multiple-choice questions (i.e., Questions 1-4) between Study 1 and Study 2. The goal was to detect to what degree the characteristics of the baseline sample extend to an exact replication with an independent sample.

RQ4: How consistent are MTurk participants when ask questions about themselves and their experiences?.

As with RQ3, this research question was evaluated using the multiple-choice questions in the survey (i.e., questions 1-4), comparing only responses of participants who completed the survey twice (i.e., groups *A* and *B* in Figure 2). The goal was to measure, at least in part, participant consistency.

3.1 Implementation

The survey was deployed as a HIT on Amazon's Mechanical Turk (MTurk). For each study, a participant could answer the survey only once. The baseline survey was available to all participants. For the second study, those who participated in the first study were blocked to enforce an independent sample of participants. In the third study, the blocked participants were unblocked, allowing participants from the first two studies to also participate in the third study.

Each study was available for two weeks. The three studies' total execution time was from August 02, 2015 - September 02, 2015, and users were paid $0.05 for each assignment. Participants had 10 minutes in which to complete all eight questions. A question could be skipped if the participant did not feel comfortable answering. We used the verification code in Question 8 to catch bots or people completing the survey haphazardly. We consider all the workers on MTurk as our study population and did not require any qualification for participation.

3.2 Analysis and Metrics

We use summary statistics to characterize the populations of each sample to answer RQ1 and a statistical test of two proportions to compare the independent samples for RQ3. We introduce consistency measures to address RQ4 and analyze the free-text responses using a bag-of-words approach to evaluate motivations for RQ2.

3.2.1 Motivations (RQ2)

We measure participant motivations concerning payment by determining if their answer for Question 6 contains any of the following keywords: "money", "cash", "income", "pay", "part time", "payment", and "earn". Those keywords are selected from the dictionary generated from a bag-of-words analysis of all responses to Question 6. For example, the response, "To earn a little extra money." indicates a participant's concern is money when choosing to use MTurk. For those participants whose motivations are not payment-related, we use the same approach to generate another dictionary of words for Question 7 and identify the common keywords by observation.

Table 1: Example for the consistency metric on the gender question

Study 1	Study 3	Label
Female	Female	consistent
Male	empty	consistent
Male	Female	inconsistent
Male	Prefer not to say	inconsistent

We did not considered the negation response, so there is a threat that some participants may have responded "not money," which would be classified as caring about payment. However, in manual inspection of some of the responses, we did not observe this.

3.2.2 Test of Two Proportions (RQ3)

A test of two proportions (i.e., `prop.test`[2] in R) can be used for testing whether the proportions in two independent groups are the same. For example, for Question 4, using participants in Study 1 and Study 2 as the independent groups, we test if the proportions of people having a CS/IT degree are the same. We find that the difference between the proportions (i.e., 109/300 for Study 1 and 66/300 for Study 2) is statistically significant with $p < 0.001$.

3.2.3 Consistency (RQ4)

We measure consistency by determining if the same worker answers consistently on the multiple-choice questions in two survey submissions. For each question, we mark their answer as consistent or inconsistent, as shown in Table 1. For example, if on Question 1, a participant claims *Male* in Study 1 and *Female* or *Prefer not to say* in Study 3, then this is marked as *inconsistent*. In the case of the same response or if one survey has a missing value, we give participants the benefit of the doubt and mark those as *consistent*. Questions 3 and 4 were evaluated using the same technique.

Regarding the age question, we allow flexibility in the second submission in the case a birthday happened between survey submissions. Thus, if a participant answered the same age in the second survey submission, or the same age plus one, it was marked as consistent.

4. RESULTS

In all, we had 1,200 submissions of the survey and did not reject any due to incorrect answers on Question 8. There were 300 participants in Study 1, 300 participants in Study 2, and 600 participants in Study 3. The overlap between Study 1 and Study 3 was 86 participants (*A* in Figure 2) and 50 participants between Study 2 and Study 3 (*B* in Figure 2), resulting in 1,064 unique participants submitting 1,200 surveys. Time to completion for Study 1 was 87 hours, 110 hours for Study 2, and 217 hours for Study 3.

Table 2 shows summary statistics for the first six survey questions for all participants in each study. The rows represent survey questions and responses and the three right-most columns present the results from each study. For example, for Question 1 and Study 1, 145 participants answered *Male*, representing 48% of the respondents. For Question 3 and

[2] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prop.test.html

Table 2: Survey questions and results. Studies 1 and 2 had 300 participants each, and Study 3 had 600 participants. All accepted responses are represented.

Question	Study 1	Study 2	Study 3
Sample size	n=300	n=300	n=600
1. What is your gender?			
- Male	48%(145)	37%(110)	43%(255)
- Female	51%(154)	63%(189)	56%(338)
- Prefer not to say	0%(0)	0%(1)	1%(4)
- Empty	0%(1)	0%(0)	1%(3)
2. What is your age?	$\mu = 34$	$\mu = 37$	$\mu = 38$
3. Which of the following best describes your education level?			
- Some High School	6%(18)	7%(20)	6%(38)
- Some college, no degree	18%(55)	26%(78)	20%(122)
- Associates degree	9%(26)	13%(38)	12%(70)
- Bachelors degree	45%(135)	36%(108)	43%(258)
- Graduate degree (Masters, Doctorate,etc.)	21%(64)	17%(51)	18%(109)
- Empty	1%(2)	2%(5)	1%(3)
4. Do you have a degree in IT, computer science, or a related field, or does your job require programming?			
- Yes	36%(109)	22%(66)	24%(145)
- No	61%(184)	77%(231)	75%(447)
- Empty	2%(7)	1%(3)	1%(8)
5. What programming languages have you used?			
- Java	17%(51)	12%(37)	16%(95)
- Html	7%(22)	5%(16)	7%(44)
- Python	2%(7)	3%(9)	5%(30)
6. Why do you use Amazon's Mechanical Turk?			
- Money	83%(249)	87%(261)	87%(524)
- No-Money	17%(51)	13%(39)	13%(76)

Study 2, 108 participants answered that they have a *Bachelors degree* (36%). Question 7 is used for other motivation identification and is discussed in Section 4.2.

4.1 RQ1: CS or IT background

To the survey question, *Do you have a degree in IT, computer science, or a related field, or does your job require programming?*, between 22% and 36% of each sample responded *Yes*. This comes from 109 participants (36%) in Study 1, 66 (22%) in Study 2, and 145 (24%) in Study 3. By combining the study responses and removing duplicates[3], the result was 246 (24%) of 1,023 unique and consistent participants answered *Yes*.

Using the same combined pool for Question 3 about education level, most participants 951(93%) have at least some college education with 717 (70%) having a college degree. A substantive number, 191 (19%), have a graduate degree.

For Question 5 about programming language experience, the top three languages are Java, HTML, and Python (note that most people without programming experience answered "none" or left this question blank). We used a bag-of-words analysis to generate a dictionary using responses to Question 5, then selecting the top words representing programming languages.

Summary.

For studies requiring a computer science background, based on our surveys, approximately 24% of MTurk worker are potentially qualified, and the most known language is *Java*. The pool of participants is also particularly well educated with 70% having a college degree.

4.2 RQ2: Motivations for Using MTurk

Echoing prior work, we found that the reason most participants (86%) use MTurk (Question 6) is for payment. For those workers whose main concern is *not* payment, they mention "time", "interesting", "type", "effort", and "difficulty" with descending frequency in Question 7.

We also analyzed the relationships of money as a motivating factor with gender, computer science background, and education level. We found no significant relationship between money and gender or money and education level. However, participants with a computer science degree (Question 4) were significantly less likely to care about money than those without a computer science degree. That is, 191 (78%) of the computer science participants care about money, versus 686 (90%) of the non-computer science participants ($p = 2.21 * 10^{-6}$ using Prop.test in R).

[3] Both surveys were removed if there were duplicates from a person who inconsistently answered at least one question, as found in RQ4

Table 3: Proportion test of independent sample from Study 1 and Study 2

Characteristics	Study 1	Study 2	p-value
1. Gender			
-Male	48%	37%	0.004934**
-Female	51%	63%	0.004934**
3. Education Level			
-Some High School	6%	7%	0.8669
-Some college	18%	26%	0.0306*
-Associates degree	9%	13%	0.1457
-Bachelors degree	45%	36%	0.0306*
-Graduate degree	21%	17%	0.2133
4. Hold CS Degree			
-CS	36%	22%	p<0.001***
-No-CS	61%	77%	p<0.001***
6. careMoney			
-Money	83%	87%	0.2128
-No-Money	17%	13%	0.2128

*$\alpha = 0.05$, **$\alpha = 0.01$, ***$\alpha = 0.001$

Summary.

Regarding concern for payment, participants with a computer science degree (Question 4) were significantly different from non-computer science participants.

4.3 RQ3: Sample Consistency

Population samples in MTurk are self-selected, which can create a bias toward people who are interested in the topic or a type of HIT. To measure the consistency between samples, we analyze the responses from Study 1 and Study 2. Sample consistency is measured on Questions 1, 3, 4, and 6 between Study 1 and Study 2, for which we obtained independent samples. We used a test of two-proportions, as described in Section 3.2, to compare the samples, as shown in Table 3.

For Question 1 and Question 4, a statistically significant difference was observed with $\alpha = 0.01$. This implies that characteristics between samples vary substantially, even for large samples with 300 participants. We note that in the study design, Study 2 was launched two weeks after Study 1, and that participants from Study 1 were excluded from Study 2. This means the population was slightly different for Study 2, which could explain differences in the population characteristics. A replication where Study 1 and Study 2 are launched simultaneously may shed some light on this observation.

Summary.

Even large, independent samples of 300 participants in the same population had significant differences in gender and CS/IT degree.

4.4 RQ4: Participant Consistency

For participants who completed two surveys, we labeled them inconsistent according to the consistency metrics described in Section 3.2. In total, there were 136 participants who completed two surveys and 41 were labeled inconsistent.

These inconsistent participants answered inconsistently on at least one question, but sometimes on three of the four questions. Table 4 summarizes the number of questions answered inconsistently between the surveys. A majority of the inconsistent participants, 36 (88%) answered just one question inconsistently, but 5 (12%) answered two or more questions inconsistently.

Table 4: Number of questions answered inconsistently (gender, age, education level, CS degree)

Questions	Participants
0	95
1	36
2	3
3	2

Breaking the data down by question, we see that Question 3 and Question 4 about education (level and computer science degrees) were the most commonly inconsistently answered, with 11%-12% of the surveys revealing inconsistencies. Table 5 presents the consistently results by survey question. Since some people answered inconsistently on multiple question, the last column, *Individuals*, summarized inconsistency by person.

As our original intention was to understand the number of MTurk participants who are potentially qualified to perform computer science related tasks, we are particularly interested in the consistency of participants with a CS degree. We observed 114 participants who answered two surveys and answered *Yes* to Question 4.[4] We divide those 114 participants into two groups by whether they have a CS degree, and call them CS group (45 participants) and Non-CS group (69 participants). This time only considering the remaining consistency questions of gender, age, and education level, we observe consistency within the CS and Non-CS groups. Table 6 shows that participants in the CS group tend to be less consistent than those who do not. In the CS group, 11 (24%) perform inconsistently, while in the Non- CS group, 9 (13%) perform inconsistently, though this difference is not statistically significant (test of two-proportions, $p = 0.1894$).

Summary.

Of the participants who answered multiple surveys, 30% answer inconsistently on at least one questions.

5. DISCUSSION

In this section, we discuss the implications of our results and threats to validity.

5.1 Implications

Our survey results show that approximately 24% of participants have a CS/IT background. Considering that there are about 500,000 workers registered on this platform [17], this supplies a considerable population who may be potentially qualified for software engineering research and evaluation, though we also recognize we survey only a small fraction of all workers.

We found most workers on MTurk are payment driven. This is consistent with findings in prior work that mone-

[4] This is different than the consistent group, since answering CS degree and empty (see Table 1) is considered consistent, but that group is omitted from this analysis since we wanted to be very sure that we only look at those with CS background.

Table 5: Answer Consistency for Repeat Participants

	Gender (Q1)	Age (Q2)	Edu Level (Q3)	CS Degree (Q4)	Individuals
Consistent	96% (131)	91% (124)	88% (120)	89% (121)	70% (95)
Inconsistent	4% (5)	9% (12)	12% (16)	11% (15)	30% (41)

Table 6: Consistency of workers with CS/IT background

	Consistent	h%	v%	Inconsistent	h%	v%
CS	34	76	36	11	24	55
Non-CS	60	87	64	9	13	45

h% for horizontal percentage; v% for vertical percentage.

tary incentive can facilitate high-quality work [21], and that many people use MTurk as a source of supplementary income [13]. However, the proportions of workers whose major concern is money between CS and Non-CS workers are statistically different. This illustrates a need to pay attention to other factors such as “time”, “type”, and “difficulty” when creating HITs to incentivize appropriate participants.

Our results for sample consistency show some significant differences, which implicate the necessity to run studies using multiple study samples in order to get access to a more representative sample population.

Our within-participant consistency analysis results show that approximately 30% of the participants answered inconsistently on at least one question. This is consistent with prior work that has found 39% of participants do not answer HITs conscientiously [3]. It implies that a non-trivial proportion of workers in MTurk try to game the system, which can have a strong impact on the outcomes of crowdsourced tasks, though we note that this work was limited to a microtask crowdsourcing environment. Still, these results point to a need for consistency checks to ensure workers are mindfully and honestly performing tasks. Related work indicates that students tend to respond conscientiously on MTurk [3].

5.2 Threats to Validity

The following threats to validity impact our results:

5.2.1 *Conclusion*

We are measuring suitability for software engineering studies based on a question about CS/IT degrees, rather than measuring related skills. Workers without degrees may be suitable for tasks, and workers with degrees may not be, depending on the required skills.

Our measurement method for Questions 5, 6, and 7 is bag-of-words, which may mis-classify negative responses (e.g., “not money”) as positive responses (e.g., “money”). While we did not observe this upon inspection of the data, in future studies, we will use multiple-select question to substitute the free-text response and facilitate a more straightforward analysis.

When measuring consistency for multiple-choice questions, in the case of missing value, participants are given the benefit of doubt and marked as consistent in order to mitigate the bias for inconsistency (Table 1). This conservative approach means our results may under-estimate the number of inconsistent participants. If we were to mark an empty response as inconsistent, the impact would have been an additional 10 inconsistent participants, meaning 51 (37.5%) would have been inconsistent.

Our study uses worker IDs, which are assigned by Amazon, to identify participants. It is possible, therefore, that multiple people could use the same MTurk account or that one person could use multiple accounts. Amazon uses reasonable measures to try to discourage these practices, but the threat remains.

5.2.2 *Internal*

For RQ4, we compared answer consistency between CS and Non-CS participants. Seven participants provided an empty answer on Question 4 for one survey, so grouping into either CS or Non-CS would be difficult. These surveys were omitted from the analysis to reduce bias.

Since we were paying participants, our samples are biased toward people who want money, impacting the results of RQ2 concerning motivations. To mitigate this threat, we paid very little ($0.05).

5.2.3 *External*

While we surveyed 1,064 unique MTurk participants, our results may not generalize to other crowdsourcing platforms. Replication with other crowdsourcing platforms, and with a larger sample, are needed to mitigate this threat.

6. CONCLUSION

Crowdsourcing platforms like MTurk provide convenient environments for conducting empirical studies. At the same time, in order to get quality experimental data, workers' consistency should not be taken for granted. In this work, we deployed three surveys on MTurk, one serving as a baseline, the second serving as a replication with an independent sample of participants, and the third as another replication with a larger sample and allowing for past participants to complete the survey again.

Our results show that a considerable percentage of workers have college degrees and are potentially qualified for software engineering related tasks. We explore participant motivations when choosing a task and find most people are money driven. Further, we compare the independent samples of 300 participants each from Study 1 and Study 2 and observe significant differences in gender and whether participants hold a CS degree. Additionally, we explore the consistency of participant answers based on different attempts for same questions, finding that a high percentage of workers (30%) perform inconsistently. These results have implications for people who run studies on MTurk, illustrating a need for large sample sizes and controls for participant quality while a study is running.

Acknowledgements

This work is supported in part by NSF SHF-EAGER-1446932 and the Harpole-Pentair endowment to Iowa State University.

7. REFERENCES

[1] R. A. Cochran, L. D'Antoni, B. Livshits, D. Molnar, and M. Veanes. Program boosting: Program synthesis via crowd-sourcing. In *Proceedings of the 42Nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '15, pages 677–688, New York, NY, USA, 2015. ACM.

[2] E. Dolstra, R. Vliegendhart, and J. Pouwelse. Crowdsourcing gui tests. In *Software Testing, Verification and Validation (ICST), 2013 IEEE Sixth International Conference on*, pages 332–341, March 2013.

[3] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: screening Mechanical Turk workers. In *Proceedings of the 28th international conference on Human factors in computing systems*, 2010.

[4] A. Finnerty, P. Kucherbaev, S. Tranquillini, and G. Convertino. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, CHItaly '13, pages 14:1–14:4, New York, NY, USA, 2013. ACM.

[5] Z. P. Fry, B. Landau, and W. Weimer. A human study of patch maintainability. In *Proceedings of the 2012 International Symposium on Software Testing and Analysis*, ISSTA 2012, pages 177–187, New York, NY, USA, 2012. ACM.

[6] Z. P. Fry and W. Weimer. A human study of fault localization accuracy. In *Proceedings of the 2010 IEEE International Conference on Software Maintenance*, ICSM '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[7] O. S. GÃşmez, N. Juristo, and S. Vegas. Understanding replication of experiments in software engineering: A classification. *Information and Software Technology*, 56(8):1033 – 1048, 2014.

[8] S.-H. Kim, H. Yun, and J. S. Yi. How to filter out random clickers in a crowdsourcing-based study? In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, BELIV '12, pages 15:1–15:7, New York, NY, USA, 2012. ACM.

[9] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[10] T. D. LaToza, W. B. Towne, C. M. Adriano, and A. van der Hoek. Microtask programming: Building software with a crowd. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 43–54, New York, NY, USA, 2014. ACM.

[11] W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

[12] M. Nebeling, M. Speicher, and M. C. Norrie. Crowdstudy: General toolkit for crowdsourced evaluation of web interfaces. In *Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, EICS '13, pages 255–264, New York, NY, USA, 2013. ACM.

[13] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[14] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *ICWSM*, 2011.

[15] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in Mechanical Turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, 2010.

[16] T. W. Schiller and M. D. Ernst. Reducing the barriers to writing verified specifications. *SIGPLAN Not.*, 47(10):95–112, Oct. 2012.

[17] N. Stewart, C. Ungemach, A. J. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, J. Chandler, et al. The average laboratory samples a population of 7,300 amazon mechanical turk workers. *Judgment and Decision Making*, 10(5):479–491, 2015.

[18] K. T. Stolee and S. Elbaum. Exploring the use of crowdsourcing to support empirical studies in software engineering. In *International Symposium on Empirical Software Engineering and Measurement*, 2010.

[19] K. T. Stolee and S. Elbaum. Refactoring pipe-like mashups for end-user programmers. In *International Conference on Software Engineering*, 2011.

[20] K. T. Stolee, S. Elbaum, and D. Dobos. Solving the search for source code. *ACM Trans. Softw. Eng. Methodol.*, 23(3):26:1–26:45, June 2014.

[21] M. Yin, Y. Chen, and Y.-A. Sun. Monetary interventions in crowdsourcing task switching. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.