SAND2014-19979C

A Study of daily sample composition on Amazon Mechanical Turk

Blind Review

No Institute Given

Abstract. Amazon Mechanical Turk has become a powerful tool for social scientists due to its inexpensiveness, ease of use, and ability to attract large numbers of workers. While the subject pool is diverse, there are numerous questions regarding the composition of the workers as a function of HIT posting time. Given the "queue" nature of HITs, and the disparity in geography, and time of HIT responders, it is natural to wonder whether HIT posting time/day can have an impact on the population that is sampled. We address this question using surveys on AMT and show (surprisingly) that there does not seem to be a statistically significant difference in terms of demographics characteristics over time.

1 Introduction

Amazon Mechanical Turk (AMT) is an online crowdsourcing platform that allows "requestors" to post tasks (called "Human Intelligence Tasks" or HITs), such as image classification and text categorization, for "workers" (or Turkers as they are called) to complete. Increasingly, AMT is being used by the academic community to conduct experiments that would normally have taken place in the lab, for instance in cognitive behavioral experiments [3], identifying subjects with particular psychiatric symptoms [17], and behavioral economics [6].

Numerous studies have shown the significant diversity of subject pools recruited from AMT [1, 3, 13, 12], thus addressing significant concerns of participant homogeniety (i.e., the WEIRDness issue [4, 16]. However, the labor market structure of AMT, and the diversity of its subject pool, may in fact cause a problem.

In AMT tasks are posted by the requestor and are available to workers for (1) a fixed, requester determined amount of time, such as 1 hour, a week, etc; (2) the requested number of workers have completed the task. Satisfying the second condition can take a variable amount of time, from minutes to weeks.

Ten's of thousands of tasks are available at any time on AMT. To manage this abundance of tasks, workers often sort the tasks by their posting date [2]; this leads to newly posted tasks showing up in the first pages of the results. The probability of a worker seeing a task that is far down in the search results may be low (perhaps following the "law of surfing" which suggests an inverse gaussian distribution for the number of items an internet user views before stopping [5, 7]).

Since tasks can quickly be pushed several pages down, it's clear that there could be a potential demographic differences to arise. That is, we can consider a task to be a sample of the workers at the time it was posted, and this sample may be biased based on the time of posting.

Some anecdotal evidence suggests this may be the case. For instance, in [9] it was found that there was a not-inconsequential difference in the gender of workers when posting a task in the morning vs. evening.

As an example, consider a restauarant in a busy section of town. If you were to sample the population at lunch time it may be filled with businesman, but at dinner time it may be filled with a completely different crowd.

Given this, our objective is to evaluate the potential impact of time of HIT posting on the demographics of the responders. This will allow us to understand potential disparities in sample composition that may influence the outcome of a task.

1.1 Survey design

We took a "panel survey" approach. We conducted 3 waves in which we posted a survey instrument (described in more detail in Section 1.2) as a HIT on AMT. Figure 1 provides a schematic of when we posted the survey instrument.

In the first wave we posted the survey instrument twice a day every day for a calendar week (7 days), starting on a Monday. We allowed 50 responses per posting of the HIT. During this week we made no limitation on the number of times participants could complete the survey (at different times). For instance, a respondent could complete the survey on Monday 8/18 at 8:00am, AND again on Monday, 8/18 in the evening. However, each respondent could only complete the survey once per HIT.

Payment was a $1.00 to the respondent. The HIT stayed open for 3 hours, although we usually had 50 responses well before then.

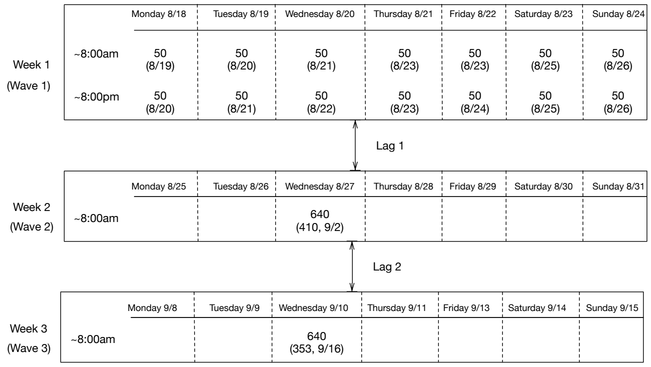
Wave 2 took place 1 week and 2 days after the start of Wave 1. We invited all participants from Wave 1 to retake the survey for the same payment. We keep Wave 2 open for a 6 days. Only one HIT was provided, and each participant could only take the survey once. Wave 3 took place two weeks after Wave 2 and 3 weeks and 2 days after Wave 1. One again, we invited all participants from Wave 1 to retake the survey. We followed the same procedure as for Wave 2.

1.2 Survey Instrument

The survey instruments consisted of approximately 120 questions that ranged from basic demographics to attitudinal, personality and psychological measures.

Of focus to this work will be the 5 demographic questions:

- In which country do you currently live? (provide list of countries)
- What age are you? (answer options will be age ranges).
- What is your household income? (answer options will be in ranges).
- What is your gender?
- What is the highest level of education completed in your household?



		Monday 8/18	Tuesday 8/19	Wednesday 8/20	Thursday 8/21	Friday 8/22	Saturday 8/23	Sunday 8/24
Week 1 (Wave 1)	~8:00am	50 (8/19)	50 (8/20)	50 (8/21)	50 (8/23)	50 (8/23)	50 (8/25)	50 (8/26)
	~8:00pm	50 (8/20)	50 (8/21)	50 (8/22)	50 (8/23)	50 (8/24)	50 (8/25)	50 (8/26)

		Monday 8/25	Tuesday 8/26	Wednesday 8/27	Thursday 8/28	Friday 8/29	Saturday 8/30	Sunday 8/31
Week 2 (Wave 2)	~8:00am			640 (410, 9/2)				

		Monday 9/8	Tuesday 9/9	Wednesday 9/10	Thursday 9/11	Friday 9/13	Saturday 9/14	Sunday 9/15
Week 3 (Wave 3)	~8:00am			640 (353, 9/16)				

Fig. 1: Schematic of survey posting. Each entry has the number of responses requested, below is the date on which the respondents were paid. For Wave 2 and 3, we received fewer than the requested responses; for Wave 1 we received all responses.

1.3 Data Collection

While HITS were used to recruit participants from AMT, the actual survey was hosted on SurveyMonkey. A "double validation code" approach was used to link AMT users to the survey response. When accepting the HIT, Turkers were given a validation code which they were to enter into the survey instrument on SurveyMonkey. At the end of the survey they were given a "survey code" to enter back into the HIT.

The main purpose of the codes were to link the respondents between AMT and SurveyMonkey without having to use respondents AMT ID (the goal being to use the AMT ID as little as possible, given it can be used to identify an individuals [10]). In addition, this verified that the Turker completed the HIT.

For this study we did not penalize the Turker if they did not correctly put in the validation or survey code.

1.4 Data Cleaning

In the following we will use this terminology:

Survey Instance A survey instance is one batch of surveys released. So it is uniquely determined by the day of the week and the time at which it was released.

Survey response (or just response) A survey response is one response by a subject to a survey. A survey response is uniquely determined by the subject id, creation date and time of the survey instances release.

Starting with 1463 responses, we removed responses where:

– The subject did not put in the correct validation code (we kept responses from individuals who put in their AMT ID for their validation code): 150

- The subjects did not correctly answer all three administrative questions: 123
- The subjects took less than 1 minutes to do the survey: 6 responses.
- The subjects did not consent to the experiment: 80 responses
- The subjects answered less than 90% of the questions: 67 responses

This left us with *1056* survey responses. Table 2 contains the number of survey responses that were kept for each wave/time period. Table 1 has the number of responses without cleaning.

Most survey instances had similar numbers of surveys discarded, so we believe there was little bias on survey removal based on time.

Table 1: Raw Counts

Wave	Count
Wave1	700
Wave2	410
Wave3	353

Table 2: Survey responses count

Wave	Time	Day	Count	Perc. Kept
Wave1	morning	Sun	42	0.840
Wave1	evening	Sun	38	0.760
Wave1	morning	Mon	31	0.620
Wave1	evening	Mon	33	0.660
Wave1	morning	Tues	33	0.660
Wave1	evening	Tues	33	0.660
Wave1	morning	Wed	33	0.660
Wave1	evening	Wed	37	0.740
Wave1	morning	Thurs	35	0.700
Wave1	evening	Thurs	34	0.680
Wave1	morning	Fri	36	0.720
Wave1	evening	Fri	36	0.720
Wave1	morning	Sat	37	0.740
Wave1	evening	Sat	36	0.720
Wave2	morning	Wed	302	0.737
Wave3	morning	Wed	260	0.737

2 Results

2.1 Overall Statistics

First we will generate the overall demographics. To do this, we took all responses from Wave 1 (we only used the first response from a respondent).

Several surveys have indicated that turkers originate from two major countries: USA and India [15, 8]. We wanted to test out the percentage of individuals who would join from the USA and India as well. To do this, one of the survey questions asked was: “In which country do you currently reside?”.

Table 3a provides the overall percentages of responses from the USA, India and every other country. Following previous work, we see that responses from the USA are prevalent, with responses from India being second. However, previous work suggested a higher response rate from India [15], which surprisingly did not appear. Note that subjects were allowed to repeat the survey multiple times during Wave 1 (see Section 1.1 for more details), however Table 3a uses the respondents first response within Wave 1.

	Country	Count	Percentage
1	USA	402	0.843
2	India	65	0.136
3	Other	10	0.021

(a) Respondent distribution over countries.

Country	<Bachelors	>=Bachelors
USA	0.47	0.53
India	0.03	0.97
Other	0.30	0.70

(b) Distribution of

Table 4: Distribution of income by country – Wave 1

	< $25,000	>= $25,0000
USA	0.19	0.81
India	0.57	0.43
Other	0.10	0.90

Figure 2 shows the education level of the subjects in Wave 1

We see similarity to the results from [15] – the overwhelming majority of Indian resident workers had a bachelors degree or higher.

In terms of gender, Table 5 provides the distribution of gender. We see more Female worker, overall, than Male workers. When dividing by country (Table 6),

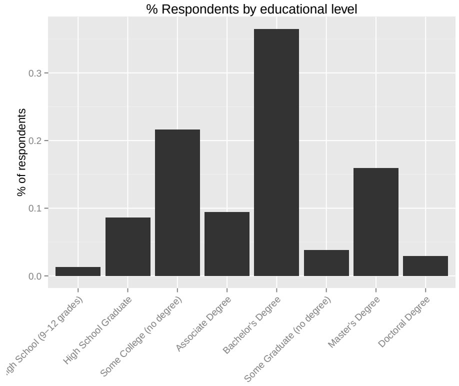


Fig. 2: Education levels for all subjects in Wave 1.

we see a large proportion of Female workers in the US, while a larger proportion of Males in India. This follows the results of [15] for their February 2010 results.

Table 5: Distribution of gender – Wave 1

Male	Female
0.4513	0.5487

Table 6: Distribution of gender by country – Wave 1

	Male	Female
USA	0.43	0.57
India	0.57	0.43
Other	0.40	0.60

Income distribution, overall, is shown in Figure 3. Table 4 provides the distribution dividing at $25,000 per country. We see the majority of US participants having an income over $25,000.

However, the question as posed was not suited for international distribution – it assumes that participants from India knew the exchange rate at the time of the survey (on 8/20/2014 it was 60 INR/ 1 USD[1])

Our overall statistics report similar statistics as [15].

[1] `http://www.x-rates.com/historical/?from=INR&amount=1.00&date=2014-08-20`

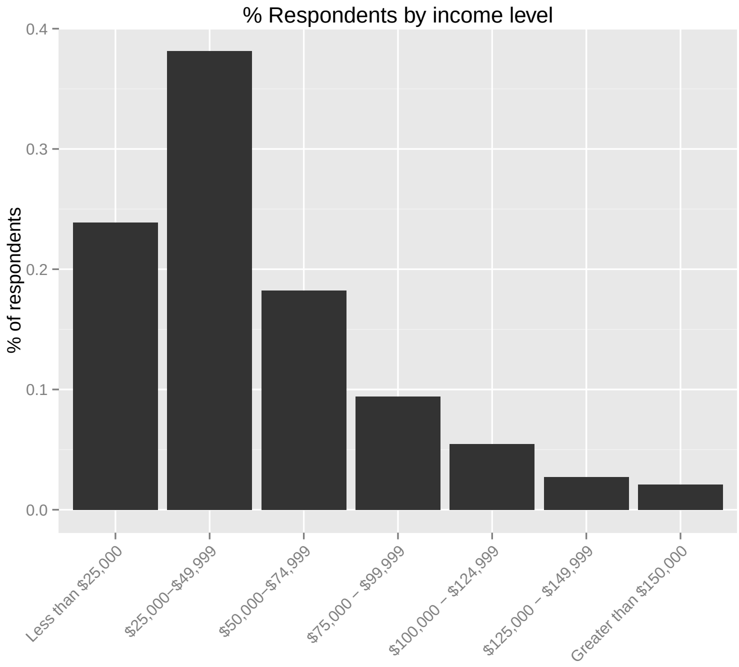


Fig. 3: Income level proportions.

2.2 Time-of-day differences

Table 7 shows distribution of demographic characteristics as a function of day of the week and time of the day that the HIT was posted. A χ^2 was performed on all demographic categories over the two categories (day of the week and time of day). In all cases except gender, no statistically significant difference was found – indicating that neither the day of the week or time-of-the day caused a change in the demographics of turkers who responded to the HIT.

Gender did have a statistically significant difference. Figure 5 shows the distribution of gender by day. 3 out of the 7 days have a majority of male respondents, whereas 4 out of the 7 days have a majority of female respondents.

3 Discussion

The results are somewhat surprising, especially for the country measure. Given the time zone difference, and the task-queue nature (as described in Section 1 we expected a significant difference in the demographics of users. Note that in several cases there seemed to be differences, but there was no statistical significance. For instance, Figure 6 seems to indicate a difference in those who respond from India on Wed.

Table 7: Demographic characteristics of participants divided by day of week and by time of day. Only Gender has a statistically significant difference over days of the week ($\chi^2 = 13.96, p = .03$ using simulated p-value with 5000 replicates)

	W1 Days							W1 Time	
	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	morning	evening
N =	(64)	(66)	(70)	(69)	(72)	(73)	(80)	(247)	(280)
Gender**									
Male	0.516	0.364	0.420	0.565	0.380	0.361	0.551	0.455	0.449
Female	0.484	0.636	0.580	0.435	0.620	0.639	0.449	0.545	0.551
Income									
<$25K	0.281	0.273	0.243	0.232	0.250	0.205	0.200	0.223	0.255
$25k-$49K	0.312	0.364	0.414	0.449	0.333	0.384	0.438	0.417	0.356
$50k-$74k	0.172	0.152	0.143	0.130	0.264	0.192	0.212	0.170	0.194
$75k-$99k	0.062	0.106	0.129	0.130	0.042	0.123	0.062	0.089	0.097
$100k-$124k	0.125	0.061	0	0.014	0.083	0.055	0.038	0.049	0.057
$125k-$149k	0.016	0.015	0.057	0.029	0	0.014	0.050	0.036	0.016
>$150k	0.031	0.030	0.014	0.014	0.028	0.027	0	0.016	0.024
Education									
Some H.S.	0	0.015	0.014	0	0.028	0.027	0	0.016	0.008
H.S. Grad.	0.062	0.045	0.057	0.072	0.125	0.096	0.112	0.089	0.077
Some College	0.234	0.242	0.243	0.188	0.181	0.178	0.212	0.190	0.231
Assoc. Degree	0.141	0.030	0.129	0.159	0.042	0.096	0.062	0.113	0.073
Bachelor's	0.344	0.409	0.371	0.348	0.389	0.356	0.375	0.360	0.381
Some Grad.	0.016	0.015	0.014	0.087	0.042	0.082	0.012	0.036	0.040
Master's	0.172	0.212	0.114	0.116	0.181	0.151	0.200	0.162	0.166
Doctoral	0.031	0.030	0.057	0.029	0.014	0.014	0.025	0.032	0.024
Country									
USA	0.797	0.788	0.829	0.783	0.875	0.877	0.838	0.846	0.810
India	0.188	0.197	0.129	0.217	0.111	0.082	0.150	0.130	0.174
Other	0.016	0.015	0.043	0	0.014	0.041	0.012	0.024	0.016
Age									
0-17	0	0	0	0	0	0	0	0	0
18-25	0.141	0.121	0.114	0.188	0.181	0.205	0.175	0.178	0.146
26-35	0.391	0.364	0.400	0.377	0.403	0.411	0.350	0.324	0.445
36-45	0.203	0.197	0.171	0.174	0.222	0.205	0.275	0.227	0.190
46-55	0.078	0.061	0.143	0.130	0.125	0.096	0.088	0.101	0.105
56-65	0.172	0.182	0.143	0.116	0.069	0.082	0.088	0.146	0.093
66+	0.016	0.076	0.029	0.014	0	0	0.025	0.024	0.020

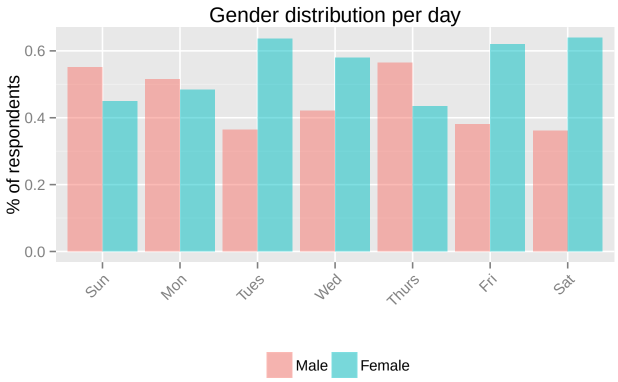


Fig. 4: Gender distribution by day in Wave 1.

Another option is that the time of acceptance of the HIT was quite different from the time the HIT was posted – thus masking the impact of time. We set the HITs to have an expiration date of 12 hours, however all of the Wave 1 HITs were completed within 6 hours, and in fact most HITs were submitted within 30 minutes of posting the HIT (median of 26.3 minutes).

One reasons may be that users are explicitly searching for certain HITs to perform – thus the impact of queue placement is mitigated. This is unlikely, as previous work ([2]) indicated that only 20% of subjects use a keyword search.

Another reason for these results may be that subjects are lying about their demographic information. This is also an unlikely possibility, as [14, 11] indicate workers self-reported demographic characteristics are stable over time. In fact, [14] conducted an independent verification of country of residence (via IP address analysis) showing that, at least for country of residence, there is a high (97.2%) match rate.

One final reason may be a bias in subject responses that are kept. There could be a correlation between demographic characteristics and subject responses which are kept – which could be masking demographic differences. Even if this were true, post-hoc data cleaning is common among AMT based studies. Thus, most studies will be analyzing data from cleaned responses anyway, and these sample composition results would hold.

A more detailed analysis of these issues is planned for future work.

4 Conclusion

The composition of the samples from AMT is an important issue that is, as of yet, understudies [12]. Through a panel survey design, we show, surprisingly, that there is no statistically significant difference in sample composition on the

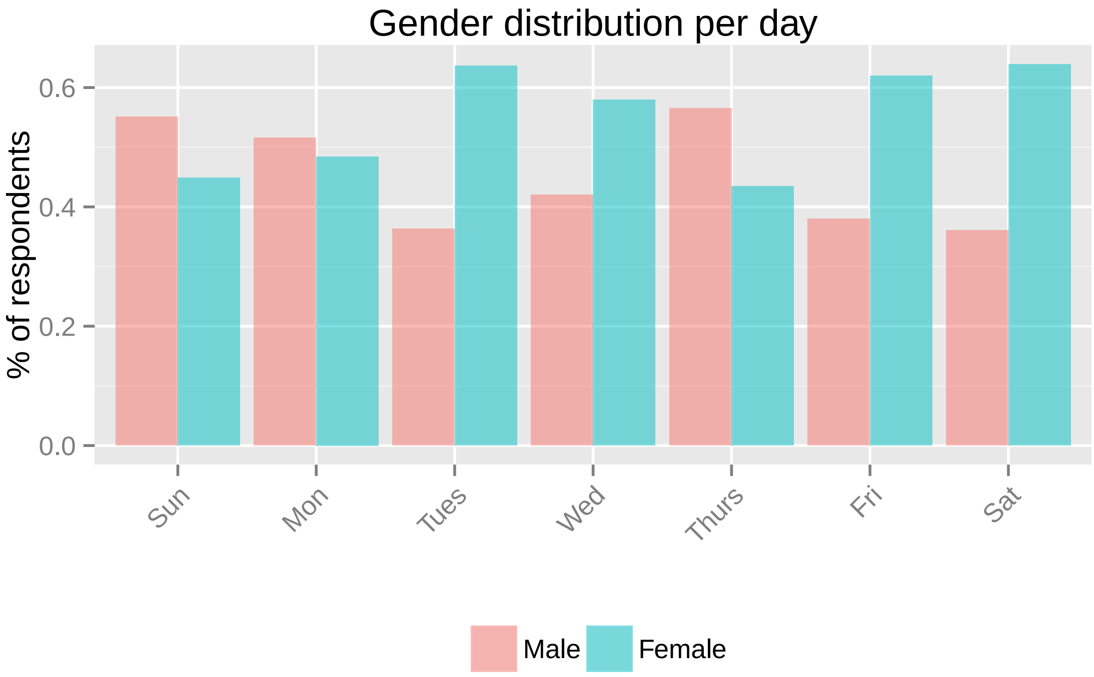


Fig. 5: Gender of respondent by day.

country of residence, income, age, and education measures as a function of when a HIT was posted (morning/evening) and day that HIT was posted (Mon-Sun).

Gender, however, does show a statistically significant difference ($\chi^2, p < .05$) as a function of day of the week. This matches results from [9] which also showed a difference in gender. However, their difference appeared as a function of the day and time of posting, whereas we were able to identify the difference as coming from the day.

References

1. Berinsky, A.J., Huber, G.A., Lenz, G.S.: Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. Political Analysis 20(3), 351–368 (Jul 2012), `http://pan.oxfordjournals.org/content/20/3/351`
2. 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. HCOMP '10, ACM, New York, NY, USA (2010), `http://doi.acm.org/10.1145/1837885.1837889`

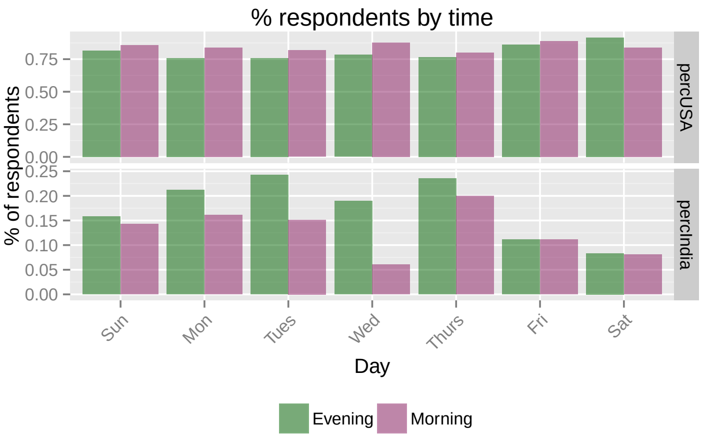


Fig. 6: Country of respondent by day, split by country.

3. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating amazon's mechanical turk as a tool for experimental behavioral research. PLoS ONE 8(3), e57410 (Mar 2013), http://dx.doi.org/10.1371/journal.pone.0057410
4. Henrich, J., Heine, S.J., Norenzayan, A.: The weirdest people in the world? The Behavioral and Brain Sciences 33(2-3), 61–83; discussion 83–135 (Jun 2010)
5. Hogg, T., Lerman, K., Smith, L.M.: Stochastic models predict user behavior in social media. HUMAN 2(1), pp. 25–39 (Sep 2013), http://ojs.scienceengineering.org/index.php/human/article/view/72
6. Horton, J.J., Rand, D.G., Zeckhauser, R.J.: The online laboratory: conducting experiments in a real labor market. Experimental Economics 14, 399–425 (2011)
7. Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M.: Strong regularities in world wide web surfing. Science 280(5360), 95–97 (Apr 1998), http://www.sciencemag.org/content/280/5360/95
8. Ipeirotis, P.G.: Demographics of mechanical turk. Tech. Rep. CeDER-10-01, New York University (2010)
9. Komarov, S., Reinecke, K., Gajos, K.Z.: Crowdsourcing performance evaluations of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 207–216. CHI '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2470654.2470684
10. Lease, M., Hullman, J., Bigham, J.P., Bernstein, M.S., Kim, J., Lasecki, W., Bakhshi, S., Mitra, T., Miller, R.C.: Mechanical turk is not anonymous. SSRN Scholarly Paper ID 2228728, Social Science Research Network, Rochester, NY (Mar 2013), http://papers.ssrn.com/abstract=2228728
11. Mason, W., Suri, S.: Conducting behavioral research on amazon's mechanical turk. Behavior Research Methods 44(1), 1–23 (Mar 2012), http://link.springer.com/article/10.3758/s13428-011-0124-6
12. Paolacci, G., Chandler, J.: Inside the turk understanding mechanical turk as a participant pool. Current Directions in Psychological Science 23(3), 184–188 (Jun 2014), http://cdp.sagepub.com/content/23/3/184

13. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on amazon mechanical turk. Judgement and Decision Making 5(5) (August 2010)
14. Rand, D.G.: The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. Journal of Theoretical Biology 299, 172–179 (Apr 2012), http://www.sciencedirect.com/science/article/pii/S0022519311001330
15. Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers?: shifting demographics in mechanical turk. In: CHI '10 Extended Abstracts on Human Factors in Computing Systems. pp. 2863–2872. CHI EA '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1753846.1753873
16. Sears, D.O.: College sophmores in the laboratory: Influence of a narrow data base on social psychology's view of human nature. Journal of Personality and Social Psychology 51(3), 515–530 (1986)
17. Shapiro, D.N., Chandler, J., Mueller, P.A.: Using mechanical turk to study clinical populations. Clinical Psychological Science 1(2), 213–220 (Apr 2013), http://cpx.sagepub.com/content/1/2/213