Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys

Frank R. Bentley
Nediyana Daskalova
Brooke White
Yahoo
Sunnyvale, CA, USA
fbentley@yahoo-inc.com
nediyana@yahoo-inc.com
brookebwhite@yahoo-inc.com

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4656-6/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3027063.3053335

Abstract
In the product design process, it is often desirable to quickly obtain information about current user behaviors for topics that cannot be obtained through existing data or instrumentation. Perhaps we would like to understand the use of products we do not have access to or perhaps the action we would like to know about (such as using a coupon) is an action taken outside of a system that can be instrumented. Traditionally, large market research surveys would be conducted to answer these questions, but often designers need answers much faster. We present a study investigating the reliability of fast survey platforms such as Amazon Mechanical Turk and Survey Monkey as compared to larger market research studies for technology behavior research and show that results can be obtained in hours for much smaller costs with accuracy within 10% of traditional larger surveys. This demonstrates that we can rely more heavily on these platforms in the product design process and provide much faster planning iterations that are informed by actual usage data.

Author Keywords
Surveys, MTurk, Crowdsourcing, Design.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

Introduction

When building new products or features, there is often a need to quickly understand existing technology use behaviors in the wider market. Often, designers or product managers will have questions on if a particular type of use "is a thing" or not in the broader population and the answers to these questions can be used to prioritize new feature/product development. Generally, rapid answers are needed to see if a particular technology behavior is something that a few people do, a good number of people do, or nearly everyone does. This data can be fairly counterintuitive from our position in Silicon Valley since we often use computing systems very differently from much of America.

Typically, in a corporate design environment, large market research surveys are commissioned to answer these types of questions. The general plan is to recruit 1,000 individuals, representative of the US population, to take a survey and answer questions about technology use or attitudes. Then data would be passed back to the product team to help with their strategic decision-making. However, this process is both very slow and expensive. Large panels, and the companies this type of work is frequently outsourced to, charge many thousands of dollars, and the turnaround times can be measured in months or quarters.

Design teams need much faster feedback if design and feature prioritization is to occur in an agile environment, with weekly cycles, as many companies currently work. For the past two years we have been experimenting with surveys on Amazon Mechanical Turk and Survey Monkey for getting very rapid feedback from a mid-size sample of distributed users for use in an agile design environment. While we were generally happy with the demographics of the users who participated in these studies, we wondered about the quality of the data we were getting from these sources. How do MTurk or Survey Monkey panels compare with the "gold standard" of a market research survey?

If we can accurately capture technology usage behaviors from these quick studies, we can change the way product planning and design works in a large company. Often, these studies can be completely run and analyzed in an hour for less than $200. Thus, a wide variety of questions that are currently answered by gut feel in the design process, due to the time and expense of collecting data, can be answered with empirical data about human behavior.

In this paper we will begin with background on survey techniques and previous research about participant demographics and data quality. We will then describe several studies that we executed simultaneously across a Market Research survey, an MTurk survey, and a Survey Monkey panel, compare the results, and discuss the relative costs of each in terms of time and money. Overall, we found very close alignment on many survey questions. We will close with implications for the product design and feature prioritization process based on our findings.

Related Work

Two popular online crowdsourcing platforms are Amazon Mechanical Turk (MTurk) and SurveyMonkey.

MTurk allows anyone to become a "requester" and post small tasks, called Human-Intelligence Tasks (HITs), which range between a few cents and a few dollars in pay. The people who complete those tasks are called "workers," and they are mostly from the US or India [9]. SurveyMonkey promises that the people who complete the surveys will be a demographically diverse population from the US, and one can pay extra for further balancing.

Large-scale market research surveys, executed by professional firms, have been considered the gold standard for survey responses, as the ideal margin of error with a sample of 1,000 out of the population of the US is 3.1%. However, these studies are costly and often take many months to complete. While surveys on other platforms such as MTurk and SurveyMonkey are affordable and allow for quick data collection, there exist two major concerns in the broader research community about using these sources: whether the survey population is representative of the general population, and whether the quality of the data is good enough to be used for design.

Previous research has shown the MTurk samples are just as demographically diverse as traditional university subject pools [9], and even more diverse than many existing Internet samples [3]. Stewart et al. showed that the diversity within MTurk samples is similar to that of the population at large [10].

Furthermore, in the domains of psychology, political science, and decision-making, research has shown that the quality of data acquired through MTurk is similar or even more representative of broader populations than data collected in more traditional ways [1, 10]. This holds for a wide variety of traditional psychology studies including Prisoner's Dilemma, the Asian disease problem [7], and the Linda Problem [8]. Horton et al. replicated the Prisoner's Dilemma experiment on MTurk and reproduced the findings from previous laboratory experiments [5]. Goodman et al. conducted a study to compare the responses of 107 MTurk participants to those of two samples (one with community members and one of students), and they found that MTurk responses are consistent with standard decision-making biases [4]. Buhrmester et al. found that a sample of 3,000 MTurk participants met the psychometric standards associated with published research [3].

Furthermore, a study by Paolacci et al. [9], compared the responses of three samples of participants: one from MTurk, one from a traditional university pool, and an online discussion board. All participants completed three classic experimental tasks: the Linda Problem, Asian disease problem, and Physician problem, and the results from the MTurk workers were similar to those of the other two samples, making MTurk a reliable source of judgment and decision-making data.

However, while standard psychology experiments have been replicated, currently the quality of data from online crowdsourcing platforms for design research in understanding technology use or preferences in digital systems is underexplored. We have not been able to find research relating to the accuracy of MTurk or SurveyMonkey panels for answering these types of behavioral technology usage questions (note that we refer to behavioral technology use, which is important for design, and not abstract behavioral studies in psychology, which have been explored on MTurk before – e.g. [6]). Thus, we set out to compare the accuracy

of responses for surveys posted to MTurk and SurveyMonkey compared to those of traditional market research surveys. Our study has great importance for being able to use and trust these sources for rapid iterations in product design or feature planning.

The Surveys

To answer our research questions around the accuracy of data obtained from MTurk or SurveyMonkey surveys, we deployed the same questions on each platform. We were able to include questions in a large Market Research survey to 1,000 respondents and ran our own surveys on Mechanical Turk (n=150) and SurveyMonkey (n=150). All data was collected in 2016.

We chose 150 users for the MTurk and Survey Monkey surveys as it represented a number that with perfect sampling could obtain a margin of error of 9% given the size of the US population. This level of accuracy is often sufficient for larger strategic or design-based questions where the desire is to understand merely if a behavior is common or if one choice is clearly better than another. A sample of 150 also provides for an affordable survey, of which dozens can be run for the cost of traditional market research, and surveys of this size could complete quickly with responses available in hours.

For our short MTurk surveys, we paid participants $0.60 for participating (an average hourly wage of $12) at a total cost of $160.50 to run, inclusive of Amazon fees. The surveys ran in about 45 minutes and results were available immediately. For Survey Monkey, we paid $1.00 per participant, or $150 overall, inclusive of Survey Monkey's fees, and the surveys took 2-3 hours to run. These are costs that most academic or industry labs can easily afford. All MTurk and Survey Monkey surveys were deployed in the evening PDT, typically around 5pm, so that participants from throughout the country could take the survey in after-work hours and we would not be biased towards missing people who worked during the day. Participants were limited to those accessing the survey from the United States, as this was the population we were most interested in studying. Exploring the use of similar survey platforms in other countries is left to future work.

On each survey platform we asked a variety of questions around technology or feature usage, all of which were relevant for various ongoing design efforts. We asked participants:

1. Do you currently have notifications enabled on your phone for your primary personal email account?
2. Have you used a coupon from your email in the past week?
3. Have you logged out of your email in the past week?
4. Have you searched your email for a package confirmation or tracking number in the past week?
5. Have you searched your email for a travel itinerary in the past week?
6. Do you have Cable Television at home?
7. Do you subscribe to Netflix?
8. Do you subscribe to Amazon Prime?
9. Do you subscribe to HBO?

	Large Survey n=1000	External Surveys	MTurk n=150	SurveyMonkey n=150
Mail Notifications On	72%	-	72% (+0%)	74% (+2%)
Coupon 1wk	46%	-	50% (+4%)	31% (-15%)
Logout 1wk	59%	-	56% (-3%)	60% (+1%)
Mail - Package	56%	-	65% (+9%)	52% (-4%)
Mail - Travel	32%	-	13% (-19%)	28% (-4%)
Cable TV	-	76%[1]	73% (-3%)	79% (+3%)
Netflix	-	65%[2]	79% (+14%)	64% (-1%)
Amazon Prime	-	34%[2]	48% (+14%)	36% (+2%)
HBO	-	15%[2]	18% (+3%)	15% (+0%)

Table 1: The percentage of respondents answering "yes" for each of the questions asked on each survey platform. A Ground Truth exists for specific sources that we did not ask about in the larger Consumer Insights survey. Conditional formatting shows discrepancies from the larger surveys, with green representing the best agreement.

Platform	Questions Answered	Age Range	Median Age	% Female	% with College Degrees	Median Income	Average Error
Large Survey	1-5	18-74	38	50%	51%	$47,500	n/a
MTurk	1	18-65	27	33%	48%	$25-50k	0%
MTurk	2	19-67	32	47%	67%	$25-50k	4%
MTurk	3	20-62	32	35%	47%	$25-50k	3%
MTurk	4-9	18-68	30	39%	46%	$25-50k	11.5%
Survey Monkey	1-5	18-78	46	53%	65%	$50-75k	5.2%
Survey Monkey	6-9	18-69	41	53%	54%	$50-75k	2.1%
US Census	n/a	0-116	37	51%	42%	$51,939	n/a

Table 2: Demographics from the various surveys with US Census numbers as a comparison. The average error observed for questions on each survey is included on the right.

Each of these questions helped a variety of teams to quickly understand market opportunities for their products and allowed designers to focus on key features with the broadest possible market appeal. We chose these items for the comparative study as they cover a variety of topics, from personal preferences in technology use (e.g., notifications and logging out), specific information-seeking behavior (searching for packages or travel), items that may be affected by income (using coupons or searching for travel), as well as content consumption questions (the variety of questions around video viewing and subscriptions). If the lower cost survey options could perform well across this set of questions, we could have confidence in using them in the future for a variety of design research questions.

Results

Table 1 compares the raw results from each of the surveys. For most question types MTurk and SurveyMonkey compared favorably to the large survey or separately obtained ground truth. When using a sample of

150, a margin of error of about 10% is expected, and most results fall in this range. For example, in the question on using mobile notifications for email, the MTurk results agree exactly with the large survey and are only 2% off from the SurveyMonkey results. Similarly accurate results were obtained for logging out (results from MTurk and Survey Monkey were within 3% of the large survey), and searching for a package (within 9%). For the media-based usage the SurveyMonkey results were within 2% of the ground truth for Netflix ,Amazon Prime, and HBO use, and within 3% on cable use, all quite good results. MTurk was much higher on Netflix (14%) and Amazon Prime (14%). The largest error was in the MTurk question around searching for flights, with an error of 19%.

The average error for MTurk across all questions compared to the larger surveys was 8% while the average error of Survey Monkey was 3.6%, showing that these rapid and inexpensive tools can be used to determine the typical levels of use of various aspects of technology. Again, as stated in the Introduction, the goal of these rapid studies is to answer in an hour or two if a certain behavior is rare, frequent, or common in the larger US population, and we were able to obtain this level of accuracy for all questions that we asked.

	Original Value	Original Error	Adjusted Value	Adjusted Error
Mail - Package	65%	9%	63%	7%
Mail - Travel	13%	19%	18%	14%
Cable TV	73%	3%	76%	0%
Netflix	79%	14%	74%	9%
Amazon Prime	48%	14%	47%	13%

Table 3: Values for MTurk responses where we observed the highest error adjusted for Income.

We will now quickly explore the demographics of participants from each of the platforms before a short discussion of some of the larger discrepancies that we observed when using these rapid survey platforms. Full demographics for each of the surveys are shown in Table 2. The larger population is generally well represented. All platforms gave us a diversity of ages, education, and income level. Most surveys received fairly good gender balancing, with a few of the MTurk studies leaning more male. Most surveys over-represented college graduates, which is likely due to the online nature of the surveys. There is no clear pattern of errors with regards to demographics, as some Turk surveys with many younger participants matched the larger population quite well while others had larger discrepancies.

Finally, we now explore some aspects that might contribute to the differences that we observed. Specifically, the largest errors were with MTurk related to the searching for travel question and use of Netflix and Amazon Prime. These larger deviations from the bigger surveys in questions 4-9 occurred with populations that are lower income which could explain fewer people checking travel notifications, fewer having

[1] http://www.pewinternet.org/2015/12/21/4-one-in-seven-americans-are-television-cord-cutters/

[2] https://blogs.adobe.com/digitalmarketing/mobile/cable-cord-consumers-loving-leaving-cheating/

cable television but more having cheaper video services such as Netflix and Amazon Prime. If we balance the results for the income distribution of the United States (see Table 3), we see the average error for these questions drops from 12% to 8.6% giving MTurk an average discrepancy over all questions of 7.1% compared to 8% without the adjustment. Other issues, such as the fact that Amazon runs MTurk and Turkers may be more likely to use other Amazon services such as Prime could account for additional discrepancies.

Discussion

We have shown that rapid online studies using MTurk or Survey Monkey samples compare favorably to larger, slower, more expensive studies. For $150 and 2 hours time, design teams can get quick answers to questions about technology use in the larger population. As is expected for a sample size of 150 people, error rates are generally within 10%, and Survey Monkey has performed slightly better than MTurk in terms of accuracy compared to larger samples from professionally-run surveys. We have shown that these online survey platforms are not just accurate representations of classic psychological test as shown in previous work [1-3, 8], but can accurately sample technology behaviors and preferences across a broader population, which is more often what we care about in design.

This type of rapid survey has the power to transform design and product management processes. We have used these methods for a wide variety of tasks over the past two years in several large product teams to make quick product or design decisions. At the most basic level, we have used MTurk to explore intelligibility of icons or choose between two use cases to prioritize. We have also used this method for higher-level strategic decisions, such as deciding to abandon a strategy for cacheing frequently viewed web pages after discovering that less than 1% of people's mobile web traffic includes revisited pages [1]. The questions on various email search behaviors (packages, coupons, and flights), helped our Mail team to prioritize new search experiences. And the question on notifications helped us in deciding to create People-only notifications[3] for Yahoo Mail.

Without quick survey methods, these questions would have taken months to answer with traditional market research methods and outsourced, expensive surveys. We frequently answer questions for teams on the same night that they have them, and this comparative work shows that these results, especially those with the Survey Monkey panel, can be trusted within about 10% to show representative behaviors.

Conclusion

We have presented a comparative analysis of rapid MTurk and Survey Monkey surveys as compared to larger, slower, and more expensive professional market research panels for answering strategic questions in the product design and feature planning process. We have shown that both Survey Monkey and MTurk can obtain results within 10% of a larger survey for a wide variety of technology usage questions.

Quick surveys can take much of the guesswork out of product planning and design and can ensure that we are building products that large numbers of people will

[3] https://yahoomail.tumblr.com/post/152342564806/an-inbox-experience-as-unique-as-you-are

have a need to engage with and can understand. From testing icons to understanding behaviors around notifications or log out frequency, getting answers from one of these new survey tools is often the fastest way to get confirmation of a design hypothesis and move on with a design using that data to guide the decision. Within hours, designers can better understand user behavior across a wide audience and better design for the population at large instead of relying on their own internal biases about what people do.

REFERENCES

1. Frank R. Bentley, S. Tejaswi Peesapati, and Karen Church. 2016. "I thought she would like to read it": Exploring Sharing Behaviors in the Context of Declining Mobile Web Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16). ACM, New York, NY, USA, 1893-1903. DOI: http://dx.doi.org/10.1145/2858036.2858056
2. Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis* 20.3 (2012): 351-368.
3. Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?. *Perspectives on psychological science* 6.1 (2011): 3-5.
4. Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26.3 (2013): 213-224.
5. Horton, J.J., Rand, D.G. & Zeckhauser, R.J. Exp Econ (2011) 14: 399.
6. Winter Mason and Siddharth Suri. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior research methods* 44.1 (2012): 1-23.
7. Amos Tversky, and Daniel Kahneman. The framing of decisions and the psychology of choice. *Environmental Impact Assessment, Technology Assessment, and Risk Analysis*. Springer Berlin Heidelberg, 1985. 107-129.
8. Amos Tversky, and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review* 90.4 (1983): 293.
9. Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5.5 (2010): 411-419.
10. Neil Stewart, Christoph Ungemach, Adam JL Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, and Jesse Chandler. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making* 10, no. 5 (2015): 479.