ASSOCIATION FOR CONSUMER RESEARCH

Labovitz School of Business & Economics, University of Minnesota Duluth, 11 E. Superior Street, Suite 210, Duluth, MN 55802

Data Collection in a Flat World: Strengths and Weaknesses of Mechanical Turk Samples

Joseph K. Goodman, Washington University in St.Louis, USA

Cynthia E. Cryder, Washington University in St.Louis, USA

Amar Cheema, University of Virginia, USA

We compare Mechanical Turk participants to community and student samples on personality, financial, and consumption dimensions, as well as classic decision-making biases. We find many similarities between Mechanical Turk participants and traditional samples, but we also find important differences that researchers should consider when using Mechanical Turk for consumer research.

[to cite]:

Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema (2012) ,"Data Collection in a Flat World: Strengths and Weaknesses of Mechanical Turk Samples", in NA - Advances in Consumer Research Volume 40, eds. Zeynep Gürhan-Canli, Cele Otnes, and Rui (Juliet) Zhu, Duluth, MN : Association for Consumer Research, Pages: 112-116.

[url]:

http://www.acrwebsite.org/volumes/1011975/volumes/v40/NA-40

Inside the Turk:
Methodological Concerns and Solutions in Mechanical Turk Experimentation

Chair: Gabriele Paolacci, Erasmus University Rotterdam, The Netherlands

SESSION OVERVIEW

Online labor markets allow "requesters" to recruit "workers" for the completion of computer-based tasks. One such market, Amazon Mechanical Turk (AMT), offers a convenient means of accessing a relatively diverse population. The speed and ease with which data can be collected on AMT has led to considerable interest in using it to collect experimental data, as indicated by the large and growing number of publications that rely on AMT data over the past few years (>400 in the social sciences alone) and self-reports by researchers subscribed to the major mailing lists in social psychology and decision-making (>50% have used AMT).

Initial evaluations of AMT as a source of data have emphasized its compelling strengths, notably the comparatively diversity of workers and the possibility of conducting research on a common population. In a nutshell, these studies have found that AMT workers produce quality data, and are more representative than other convenience samples.

Fewer efforts have been made to explore and quantify potential unique drawbacks and limitations of using AMT to collect social science data. This special session focuses on some of the issues that threaten experimental validity on AMT and on providing easily implementable solutions to avoid these problems.

The four papers included in the session deal with diverse issues. Joe Goodman discusses differences between AMT workers and more traditional subject populations that are of high relevance to consumer behavior research. Julie Downs discusses strategies for restricting data collection and data retention to attentive participants, together with their implications for the generalizability of AMT data. Dan Goldstein addresses issues of participant honesty, including the results of experiments designed to detect dishonest behaviors among AMT participants and identify some of their predictors. Gabriele Paolacci addresses the issue of non-naïvety among AMT workers by presenting studies about cross-talk and duplicate participation and provide simple remedies to attenuate this concern.

The special session contributes to the conference mission of appreciating diversity by focusing on a research method – web experimentation – that expands diversity in two important ways. First, it allows researchers to access a more representative, and certainly more heterogeneous population that traditional subject pools. Second, it democratizes science, by making these populations available to all researchers at a low cost and with minimal technical knowledge, eliminating geographic constraints and reducing financial constraints on research. Taken together, the four proposed contributions will provide attendees with a comprehensive view of how to make the best use of this resource, while avoiding common, but not widely discussed threats to data quality.

Data Collection in a Flat World: Strengths and Weaknesses of Mechanical Turk Samples

EXTENDED ABSTRACT

Mechanical Turk (AMT), an online labor system run by Amazon.com, provides quick, easy, and inexpensive access to online research participants. As use of AMT has grown, so have questions from behavioral researchers about its participants, reliability, and low compensation. A main concern about using AMT for research is that participants who are willing to participate in a study for well below the minimum wage must be unusual. And, most importantly, they might be unusual in ways that challenge the validity of research investigations. Researchers have verified that AMT demographic responses are accurate (Rand, 2012), validated the psychometric properties of AMT responses (Buhrmester, Kwang, & Gosling, 2011), and replicated some of the classic findings in behavioral economics (Horton, Rand, & Zeckhauser, 2011; Suri & Watts, 2010) and decision-making research (Paolacci. Chandler, & Ipeirotis, 2010). However, research has not thoroughly investigated differences between AMT participants and traditional samples on attention, personality, financial, and consumption dimensions. In this paper we review recent research on using AMT and compare AMT participants to community (Study 1) and student (Study 2) samples on several dimensions, finding many similarities between AMT participants and traditional samples, but also finding important differences that are relevant to consumer research.

In Study 1 we examine whether AMT participants are less attentive to study instructions than college student samples. We measured the rate at which participants pay attention by administering an attention test or Instructional Manipulation Check (IMC; Oppenheimer, Meyvis, & Davidenko, 2009) requiring careful reading of study materials. Similarly, we investigated whether AMT participants have different cognitive capabilities compared to non-AMT participants by administering the Cognitive Reflection Test (CRT; Frederick, 2005). Our results showed that AMT participants performed more poorly on the IMC and CRT compared to student participants. More importantly, we found that simply administering the attention test and filtering participants by whether they correctly answered the IMC or not reduced statistical noise; including participants who failed the IMC reduced the likelihood of finding statistically significant differences between groups on other dimensions. Though we found that IMC failure was correlated with a participant being from

outside the US or a non-native speaker of English, the results suggest the IMC was the most efficient filter as it both excluded fewer people and reduced Type II error.

For consumer behavior researchers, it is especially important to examine whether AMT participants differ in terms of how they value and spend money and time. Given that AMT participants are willing to complete tasks for little money, some have questioned their valuation of money and time. To address this issue, in Study 2 we compared AMT participants, college students and a community sample on their preferences for time and money (Cryder & Loewenstein, 2010), their material values (using the Material Values Scale; Richins, 2004), and how averse they are to spending money (using the Tightwad-Spendthrift scale; Rick, Cryder, & Loewenstein, 2008). We also compared AMT and non-AMT participants on the Big Five dimensions of personality (Gosling, Rentfrow, & Swann Jr., 2003) and global self-esteem (Robins, Hendin, & Trzesniewski, 2001).

Our results showed that compared to non-AMT participants, AMT participants were significantly (p's < .05) less extraverted, less emotionally stable, and had lower self-esteem. AMT participants also exhibited attitudes about money and time that were more similar to student participants than to community participants, valuing money more than time, reporting more materialistic values and feeling more averse to spending money (p's < .05) than the community sample. Compared to students, AMT participants were equivalent on all these dimensions. It seems that AMT participants may be similar to students in terms of their financial outlook.

Given the low compensation of AMT participants, AMT participants might also respond unusually to decision tasks involving money and risk. We explored this proposition by testing for present bias and discounting asymmetries (Loewenstein, 1988; Malkoc & Zauberman, 2006), risk aversion for gains, risk seeking for losses, and the certainty effect (Kahneman and Tversky, 1979). Our results showed that AMT participants exhibited the same effects as the student population. AMT participants were present-biased, showed delay/expedite asymmetries, were risk averse for gains, risk seeking for losses, and showed the certainty effect—but no more so than other samples.

Recent research about the use of AMT for behavioral research has concluded that AMT has many benefits, making it suitable for a wide range of behavioral research. We agree: We found that AMT participants generally produced reliable results that are consistent with previous decision making research and we found many commonalities between AMT participants and our traditional samples, contributing to this growing literature (Buhrmester et al., 2011; Paolacci et al., 2010; Rand, 2012). However, we also found important differences between AMT participants and community and student participants. To mitigate concerns that may arise from these differences, we discuss and recommend to researchers the use of screening procedures to measure participants' attention levels and acknowledge that AMT participants may vary from non-AMT participants on social and financial traits.

Screening Participants on Mechanical Turk: Techniques and Justifications

EXTENDED ABSTRACT

Concerns about the quality of Amazon Mechanical Turk (AMT) participants have led researchers to use a variety of screening techniques. Some researchers use screening data to disqualify participants in real time, punishing them for their poor performance, and some use it to omit suspicious data at the time of analysis. We assess several strategies for restricting data collection and data retention, evaluating them according to their discriminant power to identify observations contributing only noise.

We evaluated four categories of screening techniques: 1) meta-data from typical surveys, such as time on task, and depth of responses; 2) pre-screening of participants, such as limiting participation to those meeting a threshold of performance in the AMT system; 3) integral aspects of survey design, such as incentives and required responses; and 4) responses to questions included specifically to identify "poor" participants, such as gold standard questions (for which there is an objectively correct answer, unlike most survey responses). Some gold standard questions are unobtrusive, whereas others communicate their purpose to participants (e.g., by asking participants to give non-standard answers). The latter type, sometimes called Instructional Manipulation Checks (IMC; Oppenheimer, Meyvis, & Davidenko, 2009) has the potential to change participant responses in systematic ways. By randomizing whether these questions appeared at the beginning or end of the task, we created a measure of the impact of obtrusive gold standard questions on responses (conditional on passing). We included two types of criteria to assess the effectiveness of each screening technique: reliability of measures, and effect sizes of established psychological phenomena. For reliability, we used classic individual difference scales (e.g., Need for Cognition, with a Cronbach's alpha in the mid-90s; Cacioppo & Petty, 1982), as well as measures of internal consistency on behavioral tasks (e.g., choosing between lotteries that varied along dimensions of risk). For phenomenological effect sizes, we used classic demonstrations of cognitive performance and bias in judgment and decision making, such as the Stroop task and the framing effect. In addition to assessing the level of noise in the data between the screened-out and retained populations, we examined responses from these populations for evidence of systematic differences on our measures. Participants located in the US were recruited into five different surveys on AMT, each paying the equivalent of about $8 per hour, to approximate minimum wage, plus a possible bonus in some cases, to incentivize performance. For most surveys (except where noted) we required 500 or more completed AMT tasks and an approval rating of 95% or higher.

Each category of screening technique will be reviewed in turn. Simple meta-data did not prove to be useful in identifying noisy observations. The fastest 8% of respondents performed no worse than the population as a whole (N = 302, z = 0.08, p = .936), and the fastest 3% (10 of 302 participants, who performed remarkably fast with a slight discontinuity from the rest of the sample) performed only very slightly worse (z = 0.73, p =.465). The slowest 3% of the sample performed slightly better than the rest of the sample, although not significantly so. Although the faster and slower respondents did not produce notably noisier data than their more average peers, these respondents did differ along certain individual difference measures, with faster responders scoring lower in Need for Cognition and slower responders scoring higher. Removing standard pre-screening criteria did not reduce reliabilities compared to those who were required to have a large number of completed tasks and a high approval rating (N = 403; z = 0.58, p = .562). Requiring responses did not significantly improve reliabilities (N = 303; z = 1.18, p = .238). Reducing the payment rate to one-quarter of minimum wage (25¢ for an 8-minute task) had no notable effects on reliabilities (N = 346; z = 0.10, p = .920). Participants who failed gold standard questions did not perform any worse in reliabilities than those who passed (N = 178; z = 0.27, p = .787), although other differences in performance did emerge. Measures of bias did differ in some of the populations that would be omitted based on the various screeners, and obtrusive gold standard questions had moderate effects on some outcome mea-

sures, including changing the effect sizes of some measures of judgment bias. These effects will be discussed in more detail.

Consistent with other research, data quality in this sample was reasonable. Furthermore, screening strategies failed to identify meaningful subsets of the population who were contributing mere noise to the data. Although these findings cannot attest to full engagement of all participants in the tasks, they also cannot support the practice of omitting participants based on screener performance without concern about biasing the sample (at least in US populations). For example, although time stamps are a popular technique that can be used without adding to participant burden to omit people taking too little (or too much) time, these individuals' data did not warrant exclusion from analyses. Indeed, the systematic differences between high and low performance on many screening tools suggest that omitting participants based on these indicators would likely bias the sample rather than merely reduce noise. We suspect that the internal reputation system used by Amazon is effective in dissuading participants from attempting to game the system.

Under the Radar: Determinants of Honesty in an Online Labor Market

EXTENDED ABSTRACT

Many institutions and social systems depend upon some degree of honesty to function as intended. The legal system, for example, is predicated on honest testimony, and oaths are used with the goal of promoting truth-telling. Moreover, many economic transactions assume a truthful description of what is being sold or a promise that an agreement will result in a payment.

For online labor markets like AMT, honesty between the employers and employees helps the market to be efficient. Employers who trust the work of the employees, and employees who trust that payment will be rendered by the employer, both benefit from an environment in which honest dealing is the norm. Consumer behavior research, which relies heavily on self-report, is hard to verify, meaning that under prevalent dishonesty, such markets would be of limited interest to researchers.

Standard economic models capture the belief that people trade off the benefits of cheating with the costs of getting caught (Allingham & Sandmo, 1972; Becker, 1968). On AMT the costs of getting caught at any individual time are arguably low–a worker might only have their work rejected. However, consistently dishonest behavior can lead workers to be banned from the site. The frequency of cheating is an open question, determined both by the hassle of creating a new account, and recultivating the reputation necessary to complete much of the more lucrative work. Additionally, The pragmatic benefits of cheating sit in tension with people's intrinsic motivation to avoid feeling like they are dishonest (e.g., Mazar, Amir, & Ariely, 2008) and to maintain the appearance of honesty to others (Hao & Houser, 2011). Thus, it is not a priori clear how much dishonesty would be exhibited by workers in an online labor market.

The central focus of this work is measuring the degree to which workers on AMT are honest and determining which factors affect their honesty. Fischbacher and Heusi (2008) conducted a study which is the inspiration for this work. In a series of offline laboratory experiments, the authors had participants roll a die in private and report their roll. Participants were paid proportionally to the value they reported (with the exception of rolling six, which paid nothing). Since the experimenter could not see the roll, the participant could report any number. While each number would be expected 17% of the time, the subjects reported a roll of four 27% of the time and reported a roll of five 35% of the time. A roll of six, the lowest paying roll, was only reported 6.4% of the time, suggesting dishonest reporting. In addition to this baseline treatment, the authors conducted additional treatments where they increased the stakes (by a factor of three), ensured the anonymity of the participants, and changed the victim of the lie from the experimenter to another subject. These treatments did not have a large impact on the distribution of reported rolls.

In this work, we seek to understand the determinants of dishonesty in experiments in which payment can be affected by lying. We asked participants on AMT to roll die (at home or using a randomizer website, as they wished) and to report the values of the rolls, which gave them both ample opportunities to lie and no chance of being caught lying on any single roll (since we, as experimenters, could not observe the participants).

In the first experiment, we replicate the basic effect. Participants report a die roll between one and six and are paid 25 cents per pip (spot on the die). The average roll, under honest reporting, would be 3.5. The average of the reported rolls was 3.91, significantly higher than chance ($p < .0005$), with the distribution of rolls heavily favoring fives and sixes.

The second experiment asked whether people would report more honestly if there were less to be gained by lying. In the first experiment, the ratio of payouts between the worst and best roll (taking the flat fee into account) was 3.5, giving a strong incentive to cheat. In the second study, this ratio was reduced to merely 1.24. However, despite having less to gain, participants cheated as much in this condition as in the baseline, showing surprising insensitivity to what can be gained through dishonesty. Furthermore, participants from India and the US cheated by the same amount, again suggesting that the stakes are not a key determinant of cheating.

Given that cheating seems relatively unrelated to the magnitude and variance of the payouts, the third and fourth experiments ask whether the probability of detection may drive the decision to cheat. In all studies reported above, including those by other researchers, participants rolled a die just one time before reporting their answer. In such a situation, a six is just as likely as a one. However, when reporting, for example 10 rolls, it is less likely that the sum of these rolls would equal 60 by chance as it would equal, say, 35. If participants have a grasp of intuitive statistics, they would realize the experimenter could reject the null of honest reporting over multiple rolls if the sum of the die exceeds a certain number (or if the distribution of values reported deviates significantly from uniformity). In a large randomized experiment, participants rolled a die either 2, 3, 5, 10, or 20 times and were paid proportional to the sum of the result. Consistent with the view of people as intuitive statisticians, participants continued to cheat in a way that was easily detectable at the aggregate level, but undetectable at the individual level.

Non-naïvety Among Experimental Participants on Amazon Mechanical Turk

EXTENDED ABSTRACT

Certain experimental paradigms strongly rely on participant naïvety, either as a precondition for an effect to emerge, or to prevent experimental demand effects. Prior knowledge about the purpose of an experiment, familiarity with an experimental manipulation, or reason to suspect deception, can influence participant responses. While traditional subject pools offer a continuous supply of naïve participants, this is less true of AMT, where workers can complete an unlimited number of experiments. Given the popularity of AMT among consumer researchers, it is important to know whether concerns of non- naïvety among AMT workers are negligible or not.

Further, if non- naïvety is prevalent, it is pressing to come up with solutions that can be implemented by single researchers or scientific communities to mitigate this problem. In this work, we discuss two potential sources of non-naivety on AMT.

One important phenomenon that affects participant naïvety is cross-talk between participants. Empirical research on college undergraduate populations has demonstrated that participants do share information with each other, at least when sufficiently motivated (e.g., when incentives are offered for a correct response; Edlund, et al., 2009). The web offers great opportunities for cross-talk: Indeed, AMT workers maintain online forums where they share information and opinions, which could potentially lead to foreknowledge in experimental participants.

A second concern is participation in experiments that share independent or dependent measures. AMT automatically prevents workers from completing a single task multiple times. However, it is still possible that participants are recruited for experiments that are conceptually or methodologically related to experiments they have previously completed. Our survey data (Study 1, N = 300) show that cross-talk is not a critical issue on AMT. Only 26% of participants reported knowing personally someone else who used AMT, and only 28% reported reading forums and blogs about AMT. Further, when asked to rank the reasons why they discuss or read about AMT, the actual purpose or contents of the tasks are far less important than pragmatic considerations such as the amount requesters pay or their reputation. Only half of the respondents who actually read blogs (about 13% of the overall sample) reported ever seeing a discussion about the contents of a social science research study online. Such low levels of reported cross-talk can hardly contribute substantially to participant non- naïvety. However, researchers should probably monitor discussion boards that refer a lot of respondents, and conclude their experiments by asking workers how they found the task. We also highlight the less tangible effects of workers discussing the reputations of individual experimenters and research institutions.

Duplicate participation is a more serious concern. 55% of our worker sample from Study 1 reported having a list of favorite requesters and monitoring their tasks (indeed, browser plug-ins are available that do this automatically), and 58% of the time this list included academic researchers. Data from a follow up conducted one year later showed that these percentages became 63% and 71% respectively, suggesting that this is a growing concern. Moreover, a substantial proportion of workers reported participating in some of the more common and easily describable experimental paradigms (e.g., 52% of Study 1 participants played an Ultimatum Game, becoming 83% one year later).

In order to obtain more reliable information about duplicate participation, we pooled the data from several researchers who had received a total of 16,408 completed submissions (Study 2). The tasks included in this sample had been completed by a total of 7,498 workers. The average worker completed more than two studies, with the most prolific 10% of the workers (N = 750) responsible for 41% (N = 5,864) of the completed submissions. Taken together, these results suggest that duplicate participation is a potential source of non-naïvety that cannot be neglected by researchers who use AMT. At a minimum, experimenters should ask participants whether they have completed similar experiments before and treat prior participation as an additional factor in their data analysis. We offer several practical solutions that allow duplicate workers to be filtered out before they participate, saving money, and eliminating the concern that excluding duplicate workers may contribute additional researcher degrees of freedom.

The very features that make online labor markets appealing to researchers, such as its accessibility, lead to some concerns about whether experimental participants are naïve enough to participate in all experiments. Whereas participant cross-talk seems not to constitute a problem, care is required to deal with duplicate participants.

REFERENCES

Allingham, M. G., & Sandmo, A. (1972), "Income tax evasion: A theoretical analysis," *Journal of Public Economics*, 1, 323-338.

Becker, G. (1968), "Crime and punishment: An economic approach," *Journal of Political Economy*, 76(2), 169-217.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011), "Amazon's Mechanical Turk: A new source of cheap, yet high-quality, data?," *Perspectives on Psychological Science, 6,* 3-5.

Cacioppo, J. T. & Petty, R. E. (1982), "The Need for Cognition," *Journal of Personality and Social Psychology*, 42 (January), 116-31.

Cryder, C., & Loewenstein, G. (2010), "The time versus money scale," Unpublished data*,* Olin Business School, Washington University in St. Louis.

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

Fischbacher, U., & Heusi, F. (2008), "Lies in disguise: An experimental study on cheating (Research Paper Series No. 40)," *Thurgau Institute of Economics and Department of Economics at the University of Konstanz*.

Frederick, S. (2005), "Cognitive reflection and decision making," *Journal of Economic Perspectives, 19*, 25-42.

Gosling, S.D., Rentfrow, P.J., & Swann Jr., W.B. (2003), "A very brief measure of the big-five personality domains," *Journal of Research in Personality*, *37*, 504-528.

Hao, L., & Houser, D. (2011), "Honest lies," *Discussion Paper, Interdisciplinary Center for Economic Science, George Mason University*.

Horton J.J., Rand D.G., & Zeckhauser R.J. (2011), "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics*, 14, 399-425.

Loewenstein, G. (1988), "Frames of mind in intertemporal choice," *Management Science*, *34*, 200-214.

Malkoc, S. A., & Zauberman G. (2006), "Deferring versus expediting consumption: The effect of outcome concreteness on sensitivity to time horizon," *Journal of Marketing Research, 43*, 618-627.

Mazar, N., Amir, O., & Ariely, D. (2008), "The dishonesty of honest people: A theory of self-concept maintenance," *Journal of Marketing Research*, *45*, 633-644.

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

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010), "Running experiments on Amazon Mechanical Turk," *Judgment and Decision Making, 5,* 411-419.

Rand, D.G. (2012), "The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments," *Journal of Theoretical Biology*, *299*(4), 172-179.

Richins, M. L. (2004), "The material values scale: Measurement properties and development of a short form," *Journal of Consumer Research*, *31*, 209-219.

Rick, S. I., Cryder, C. E., & Loewenstein, G. (2008), "Tightwads and spendthrifts," *Journal of Consumer Research, 34*, 767-782.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001), "Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale," *Personality and Social Psychology Bulletin, 27,* 151-161.

Suri, S., & Watts, D. J. (2011), "Cooperation and contagion in webbased, networked public goods experiments," *PLoS One, 6*(3).