Towards a Guideline for Conducting Economic Experiments on Amazon's Mechanical Turk

Tim Straub[a], Florian Hawlitschek[a], Christof Weinhardt[a]

[a]*Karlsruhe Institute of Technology*

Abstract

During the last decade Amazon Mechanical Turk has evolved to an established platform for conducting behavioral research. However, designing and conducting economic experiments on online labor markets remains a complex and ambitious task. In comparison to laboratory environments, a set of specific challenges (such as synchronization and control) has to be thoroughly addressed. In order to support researchers in fulfilling this task, we provide a framework of a continuously updating guideline for conducting economic experiments on Amazon's Mechanical Turk. Our main contributions are the proposition of (i) a challenge-oriented view based on common experimental economics practices, and (ii) a collaborative continuous guideline approach.

Keywords: Crowd Services, Crowd Work, Crowdsourcing, Guideline, Online Experiments, Behavioral Experiments, Experimental Economics

1. Introduction

In recent years, researchers from various fields have recognized and started to harness the potential of conducting experiments on crowd work platforms such as Amazon Mechanical Turk (MTurk). The validity of such experiments

Email addresses: tim.straub@kit.edu (Tim Straub), florian.hawlitschek@kit.edu (Florian Hawlitschek), weinhardt@kit.edu (Christof Weinhardt)

in comparison to economic laboratory experiments was comprehensively discussed in the literature (Paolacci et al., 2010; Chilton et al., 2010). In recent years a variety of experiments was conducted on MTurk and in fact the number is still growing fast. Reasons for that are inter alia the large subject pool and low cost labor (Mason and Suri, 2012; Paolacci et al., 2010). However, compared to classical laboratory experiments, researchers are facing several challenges while conducting experiments in the crowd, e.g. synchronization and control. To overcome such issues on crowd work platforms, guidelines leading through the process of conducting experiments are needed.

A set of comprehensive guidelines for behavioral research on crowd work platforms has already been published (Mason and Suri, 2012; Horton et al., 2011; Paolacci et al., 2010). However, these guidelines often do not focus on economic experiments but on behavioral research in general. Furthermore, against the background of the fast growing number of experiments, especially on MTurk, workers start to get used to certain experiment types (Chandler et al., 2014). As a consequence, guideline-based experimental design approaches become outdated rapidly. Solutions to overcome this issue need to be developed steadily since researchers have to address the new insights of crowd workers. This stresses the need for a continuously updating guideline with a specific focus on economic experiments.

This paper proposes the outline of a state-of-the-art guideline for crowd experiments including common challenges and best practices from the experimental economics literature. We use the work of Friedman and Sunder (1994), depicted in figure 1 as a starting point and transfer it to a conceptual framework. To this end, the experimental design stage is used to exemplify our concept. Furthermore we propose an architecture of an open platform facilitating continuous updates of the guideline. Therefore, section 2 introduces the conceptual framework of a new guideline structure based on state-of-the-art literature on experiments in crowdsourcing environments. Section 3, includes a proposal for an architecture that enables researchers to collaboratively build a continuously updating guideline. Section 4 summa-

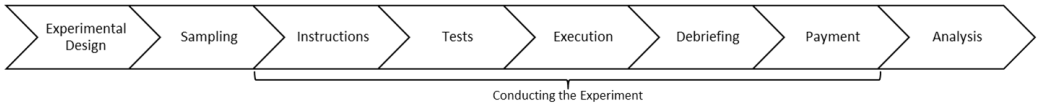

Figure 1: Process of conducting experiments based on Friedman and Sunder (1994).

rizes this work and outlines the next steps for future research.

2. Guideline for Designing a Crowd Work Experiment

We base our guideline concept on an established process from the experimental economics literature (Friedman and Sunder, 1994) as depicted in figure 1. The process can be divided in (i) the experimental design stage focusing on which experimental setup suits the research question best, (ii) the sampling or recruitment of subjects, conducting the experiment, and (iii) the analysis of the results. When transferring these steps to crowd experiments, researchers have to address certain challenges to secure result quality. In our guideline approach we highlight these challenges and suggest possible solutions retrieved from literature. In the following we present an example of our approach for (i) the experimental design stage.

2.1. Outline of a Guideline for the Experimental Design Stage

One of the first steps in experimental research is to decide which experimental design suits the research question best. Common design decisions comprise whether an experiment requires (a) asynchronous or synchronized decision-making and if it should be conducted as (b) a field or a laboratory study (Friedman and Sunder, 1994).

(a) In asynchronous as opposed to synchronous experiments subjects do not compete simultaneously against each other. Since in laboratory studies subjects usually are in the same room during the observation, it is easy to implement both setups. On MTurk however, the implementation of synchronous experiments is challenging. Usually tasks on MTurk are designed as open calls (Howe, 2006). It is unclear if or when a worker starts completing the task. This makes it difficult to realize experiments where two or more participants have to compete simultaneously against each other due to unknown arrival times of workers (Mason and Suri, 2012; Mao et al., 2012).

Challenge 1: Synchronization and arrival times.

First, it should be checked, whether the underlying research question can be addressed with an asynchronous experiment design as well. If possible, the experiment can be redesigned in an asynchronous setup. Second, if subjects

do not have to compete live against each other, playing against historical data from an earlier observation is possible (Amir et al., 2012; Suri and Watts, 2011; Straub et al., 2015). Third, if the live interaction and reaction between subjects is indispensable a waiting room can be implemented (Mao et al., 2012; Mason and Suri, 2012; Paolacci et al., 2010). However, a shortcoming of this approach is that it is unclear how long a subject has to wait, since arrival times vary. Paying a fixed fee after a certain amount of waiting time or using bots in case that waiting times are too long are possible approaches to address this (Horton et al., 2011).

(b) In a field experiment subjects usually do not know that they are being observed and the experimental setup is camouflaged. This leads to a high external validity but lower internal validity. In Laboratory studies the actions of a subject are observed in a controlled environment - the laboratory. Isolated cabins prevent unregulated contact and ensure that subjects are not influenced by uncontrolled stimuli. Variance in noise, light, and technical factors like input devices, monitors, etc. can be prevented. Therefore external confounding factors are minimized and internal validity is higher (Friedman and Sunder, 1994). In theory both, field and laboratory setups, can be realized on MTurk. However, the internal validity of experiments on MTurk compared to laboratory settings might be lower due to less possible control. To be more specific, workers might not pay attention during the observation (Paolacci et al., 2010; Horton et al., 2011; Mason and Suri, 2012). Since subjects on MTurk usually work from their own computer it is impossible to control their environment during the observation (Chandler et al., 2014; Crump et al., 2013; Rand, 2012). Chandler et al. (2014) find that subjects are watching TV or listening to music while working on MTurk. Another problem with crowd work is that some of the workers try to maximize their payout by finishing as many tasks as possible, often just clicking through them. So called malicious workers or "spammers" are not paying attention and jeopardize the overall data quality of results.

Challenge 2: Control and attention.

First, the overall task design can be aligned to incite workers to take the task seriously. Tasks that are fun, interesting, and meaningful incite subjects to pay attention (Kittur et al., 2013; Crump et al., 2013). Layman and Sigurdsson (2013) show that tasks designed as a game are more satisfying for a subject and thereby motivate to pay attention. Furthermore, researchers

could state the expected result quality and give context about the overarching goal in the instructions to give the task a meaning (Oh and Wang, 2012; Oppenheimer et al., 2009). Stating that the task is an experiment and participation helps research can as well give the task a meaning, if experimenter bias is not a problem (Orne, 1962). Second, besides redesigning the overall experimental task, experimenters can try to exclude subjects who do not pay attention before the actual observation. Many researchers test if a subject is paying attention during the instructions and exclude those who fail the respective test from the sample (Paolacci et al., 2010; Peer et al., 2013; Oppenheimer et al., 2009; Paolacci and Chandler, 2014). Oppenheimer et al. (2009) introduced the instruction manipulation check, which was recently applied by many researchers (Straub et al., 2015; Hall and Caton, 2014). The fundamental idea of the instruction manipulation check is to trick inattentive subjects with a question or free text field which is easy and at best straightforward to answer, e.g. "What is your age?". The instructions state at one point that this particular question should be ignored. Consequently subjects who do not read the instructions carefully, e.g. stating their age, can be excluded from the task (Goodman et al., 2012).

3. Framework for a Collaborative and Continuous Guideline

A continuously updating guideline can be successfully put into practice within a collaborative process. We propose that researchers should work together to integrate their (new) insights to an online platform. The proposed framework is structured as follows: First, if researchers decide to conduct a crowd experiment they can retrieve the most recent version of the guideline from the platform (as depicted in figure 2 of the appendix). Second, challenges that apply to the experimental setup can be identified and solutions can be found in the registered insights from other researchers. Third, researchers can either incorporate these solutions to their experimental setup or develop new solutions based on the challenges and hints from the platform. Fourth, after the researchers conducted their experiment they can update the platform based on their findings, e.g. if and how good the applied solutions worked. Through this process continuous updates to the guideline are facilitated based on collaborative input from researchers and the community. However, certain requirements (req.) should be implemented. Malicious workers might try to trick the system by accessing the platform in order to

get defective insights. Therefore access should be restricted to researchers. Login systems only giving access to "edu" domains or account confirmation must be implemented.

Req. 1: Access should only be given to researchers.

Overall the most important factor for a continuously updating guideline is the integration of new insights and results. Therefore researchers must be incited or enforced to incorporate their knowledge to the platform. Social rankings raising reputation and chances for citations might incite participation. Another conceivable way would be to enforce participation by restricting access with a fee, which a researcher gets back once he updates the platform.

Req. 2: Incentives or enforcements to update the platform are needed.

To secure an overall style of the guideline design principles must be set. Overarching categories with examples should be derived to make the platform easy to use and accessible for researchers.

Req. 3: Design principles should be derived for updating the platform.

4. Summary and Outlook

In this paper we proposed a guideline structure with a challenge and solution oriented view on conducting behavioral experiments on crowd work platforms. Such a guideline gives behavioral economists an easy introduction to crowd experiments with a clear and familiar structure. Furthermore we proposed a framework for an online platform where researchers could collaboratively and continuously update such a guideline. Both guideline and platform are currently in a conceptual stage. Following the notion of collaborative work, crowd work, and open innovation we plan to develop a demonstration version to incorporate practitioners and community feedback in future iterations. Hence, the next steps of future work comprise elaborating and finalizing the guideline concept and integrating it in a platform to facilitate a live deployment. Future work should analyze how researchers could be motivated to participate and how the platform could be extended. Possible extensions include experimental databases to look up which experiments other researchers already conducted and worker databases to block users who already participated in similar experiments as proposed by Chandler et al. (2014).

Acknowledgements
We would like to thank Sarah Kexel for her active support and contributions to the development of this guideline concept.

Appendix

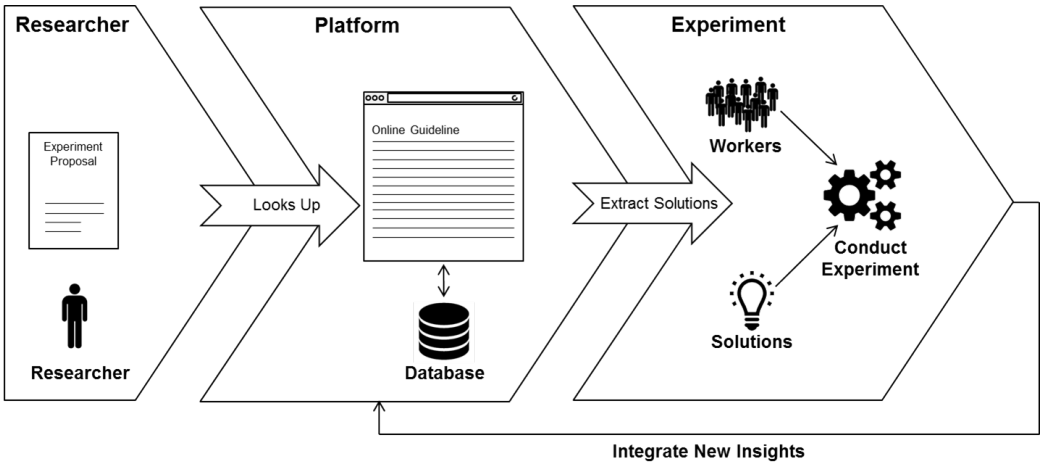

Figure 2: Framework of an online platform for a collaborative and continuous guideline.

References

Amir, O., Rand, D. G., and Gal, Y. K. (2012). Economic games on the internet: The effect of $1 stakes. *PLoS ONE*, 7(1):31461.

Chandler, J., Mueller, P., and Paolacci, G. (2014). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods*, 46(1):112–130.

Chilton, L. B., Horton, J. J., Miller, R. C., and Azenkot, S. (2010). Task search in a human computation market. In *Proceedings of the ACM SIGKDD workshop on human computation. ACM*, pages 1–9.

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE*, 8(3):e57410.

Friedman, D. and Sunder, S. (1994). *Experimental Methods: A Primer for Economists*. Cambridge University Press.

Goodman, J. K., Cryder, C. E., and Cheema, A. (2012). Data collection in a flat world: The strenghts and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224.

Hall, M. and Caton, S. (2014). A crowdsourcing approach to identify common method bias and self-representation. In *IPP2014: Crowdsourcing for Politics and Policy*.

Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, 14(6):1–4.

Kittur, A., Nickerson, J. V., Bernstein, M. S., Gerber, E. M., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. J. (2013). The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work. ACM*, pages 1301–1318.

Layman, L. and Sigurdsson, G. (2013). Using amazon's mechanical turk for user studies: Eight things you need to know. In *Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on. IEEE*, pages 275–278.

Mao, A., Chen, Y., Gajos, K. Z., Parkes, D., and Procaccia, A. D. (2012). Turkserver: Enabling synchronous and longitudinal online experiments. In *Proceedings of the Fourth Workshop on Human Computation (HCOMP'12). AAAI Press*.

Mason, W. and Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavioral Research Methods*, 44(1):1–23.

Oh, J. and Wang, G. (2012). Evaluating crowdsourcing through amazon mechanical turk as a technique for conducting music perception experiments. In *Proceedings of the 12th International Conference on Music Perception and Cognition*, pages 1–6.

Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872.

Orne, M. T. (1962). On the social psychology experiment: With particular reference to demand characteristics and their impications. *American psychologist*, 17(11):776.

Paolacci, G. and Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Curent Directions in Psychological Science*, 23(3):184–188.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgement and Decision Making*, 5(5):411–419.

Peer, E., Vosgrau, J., and Acquisti, A. (2013). Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavioral Research Methods*, 46(4):1023–1031.

Rand, D. G. (2012). The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299:172–179.

Straub, T., Gimpel, H., Teschner, F., and Weinhardt, C. (2015). How (not) to incent crowd workers. *Business Information Systems Engineering*, 57(3):167–179.

Suri, S. and Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE*, 6(3):e16836.