Conducting Large-Scale Online Experiments on a Crowdsourcing Platform

Felix Bader[1,*] and Marc Keuschnigg[2]

[1]School of Social Sciences, University of Mannheim, A5 6, 68159 Mannheim, Germany
[2]Institute for Analytical Sociology, Linköping University, Norra Grytsgatan 10, 601 74 Norrköping, Sweden. * Corresponding author; bader@uni-mannheim.de

Published January 2018 in *SAGE Research Methods Cases*

Crowdsourced online experiments are a promising complement to laboratory research in the social sciences. Conducting experiments over the Internet is not just a quick and inexpensive way of data collection. Transporting a homogeneous decision situation into heterogeneous living conditions, online experiments can generate behavioral data from people with different social backgrounds. If paired with recruiting from a global crowdsourcing platform, the potential for participant heterogeneity rises, including different nationalities and cultures. We guide through the 10 steps necessary for the practical implementation of a crowdsourced online experiment. For illustration, we refer to our recent project on fairness behavior conducted over Amazon Mechanical Turk. We used the platform to recruit participants from the United States and India. We address general issues of experimental research including randomization, instructions, incentives, and participants' informed consent as well as specific issues of online implementation such as international recruiting, payment, field control, and matching of asynchronously participating subjects. Our exemplary project focuses on altruism, fairness, and costly punishment as measured in standard experimental games. First, we randomly assigned participants to different levels of monetary incentives. We found that monetary incentives induce more selfish behavior, but the exact size of the stake is irrelevant for observed rates of prosocial behavior. Second, we explored context effects on elicited behavior using variation in participants' geographical location. We showed that context effects of regional prosperity and local social capital are comparable in size to stake effects. More importantly, we demonstrated that context effects are visible and quantifiable in a large-scale online experiment. We thus argue that studying the diverse backgrounds people bring into the experiment is the key potential of crowdsourced online designs.

Learning Outcomes

By the end of this case, students should have

- understanding of the practical steps necessary to undertake a crowdsourced online experiment,
- ability to adapt the criteria of successful laboratory experimentation to evaluate the quality and rigor of an online experiment,
- awareness of the practical challenges in operating a research project online with physically and culturally distant participants,
- knowledge of an exemplary application that brings regional context into sociological experiments.

Case Study

Project Overview and Context

In our project—published 2016 in the journal *Social Science Research*—we investigated how monetary incentives and the background characteristics people bring into the experiment influence fairness behavior. We recruited a diverse participant pool from the United States and India ($N = 991$), using the global crowdsourcing platform Amazon Mechanical Turk (MTurk). We then redirected crowdworkers to take the experiment on our own website. Data collection ran from September 4 to December 13, 2014. Participants played Dictator (DG) and Ultimatum Games (UG), which we used to elicit expectations about social norms in different regional contexts. Because socially acceptable behaviors diverge from individual motivations, social scientists use these experimental games to measure prosocial behavior and norm adherence:

- In the DG, a participant receives a monetary stake and can decide on how much of the pie (0%–100%) he or she passes to a passive receiver. Giving is typically interpreted as a manifestation of prosocial preferences.
- In the UG, a participant receives a stake and can decide on how much of the pie (0%–100%) he or she offers to a responder. The responder can then accept (and both receive their share) or decline the offer (and both receive nothing). Giving reflects both prosocial preferences and fear of sanction.

In our experiment, monetary stakes made dictators more selfish compared to a treatment with hypothetical stakes, but the size of positive stakes did not make a difference. For proposers in the UG, stake variations had no influence on decision behavior whatsoever.

Our second idea was to consider participants' social context and its potential correlation with observed rates of behavior. The countries we recruited from provide large numbers of crowdworkers at MTurk and, most interestingly for our study, are culturally distant from

each other while being heterogeneous within. We found, for example, that people from both India and the American Deep South share less. In contrast, U.S. participants from regions with high average income or high social capital are particularly generous. Our project demonstrated that social context matters for fairness behavior and that context effects—which are mostly ignored in the experimental literature—are visible and quantifiable in online experiments.

Although we had conducted experiments in brick-and-mortar laboratories earlier, this project was our first implementation of an experiment over the Internet. Starting the project in 2014, we had little practical guidance and materials available on how to set up and maintain a large-scale online experiment. Having gone through a process of learning-by-doing, we now want to share our experiences and make it easier for others to start experimenting online. This method case describes how to implement a successful crowdsourced online experiment and solve a number of practical problems that may arise while setting up and maintaining it.

Crowdsourced Online Experiments

Experiments in general consist of a treatment manipulated by the researcher. Participants are randomly assigned to receive the treatment or not (control group). As there is no self-selection into the treatment (a major problem with observational data), participants' characteristics (even their unknown features) are on average equal in treatment and control groups. Hence, any observed differences in average behavior between treatment and control groups must be caused by the treatment. This logic summarizes the power of controlled randomization for drawing causal inferences (internal validity).

Most experiments are conducted in social science laboratories, where the experimenter can further control the surroundings of the decision situation to isolate the treatment effect. The main lines of critique against laboratory experiments are the artificiality of the decision situation and the recruitment of easy-to-reach participants. Such convenience samples typically contain locally recruited students from universities in North America and Europe.

Crowdsourced recruiting for online experiments, on the other hand, allows the sampling of more diverse participants with regard to sociodemographic background, regional context, and nationality. It makes the tedious build-up of a local subject pool and the costly maintenance of a physical laboratory obsolete. In addition, conducting an experiment online typically goes along with lower participant compensation, as participation effort decreases (e.g., no transportation costs to get to the lab).

Besides saving time and money, there are other, better arguments in favor of online experiments as a promising complement to experimental data collection in the social sciences: Every physical lab is different. Chairs, windows, daylight, the position of and the distance to the experimenter, and experimenter characteristics vary between labs and could potentially influence targeted behavior. These potential confounders limit the comparability of results attained in different labs. In online experiments, there are no effects from local laboratories or experimenters, as participation occurs in an identical online environment. Individual circumstances of participation may differ, but they do so randomly and not in a systematic way.

Electronically mediated interaction further increases anonymity both between participants and toward the experimenter. Most importantly for our application, online experiments have the potential to transport a homogeneous decision situation into heterogeneous living conditions. To fully benefit from this technical advantage, online experiments can be combined with recruiting from global crowdsourcing platforms. In our project, we used MTurk where

- interested individuals can register as workers,
- organizations or individuals wanting to get work done can register as requesters,
- requesters prepare and upload work packages called Human Intelligence Tasks (HITs),
- workers see a list of available HITs showing their title, payment, and deadline,
- workers decide when and how much they work on which HIT,
- upon accepting a HIT, workers either access instructions and complete the task directly on the platform or, typical for academic studies, the HIT redirects workers to another website,
- after completing a HIT, workers submit their results,
- if the requester approves the service, workers receive payment through their Amazon accounts,
- requesters pay workers' earnings plus a 20% commission to Amazon.

Currently, MTurk sustains a large and heterogeneous participant pool of over half a million registered workers from 190 countries. Of these, roughly 75% are from the United States and 15% from India. The website MTurk Tracker (see Ipeirotis, 2010 in our list of web resources) provides a daily updated description of the participant pool. Most workers are non-students, many work on a regular job, and most are older than the student participants in traditional university subject pools. One can safely conclude that this worker population is heterogeneous with regard to sociodemographic and regional background.

MTurk was not created for academic research, but can be used for it. Typical HITs request workers to categorize pictures and videos or to transcribe audio files. Many argue that MTurk is a real labor market, in which requesters post hard-to-automate tasks and workers seek to maximize their profits (see, particularly, Horton, Rand, & Zeckhauser, 2011). In addition, online surroundings at MTurk may be less artificial and offer more familiar testing conditions than what is provided for in traditional laboratory experiments. Besides these apparent benefits, crowdsourced online experiments also share some drawbacks:

- In contrast to laboratory implementations, the online experimenter has no direct control over participants' social or physical surroundings.
- It is tough to obtain real-time interaction in an online session, as potential participants log in at separate times and may drop out along the way.

- Many workers, especially from the United States, are experienced study participants and thus can be considered non-naïve subjects. Experience with social science research may influence elicited behavior and the specific survey data you want to collect (see, for example, Chandler, Mueller, & Paolacci, 2014; Rand et al., 2014 for discussions).

We will address these limitations, among other things, in the practical implementation section*.

Practical Implementation

The following 10 steps explain the practical implementation of a successful online experiment. Many steps equally apply to the online collection of survey data over MTurk. Our description follows a participant's path through the experiment. For brevity, we exclude other relevant aspects of conducting (web-based) research, such as finding a research question and respective treatments, web programming, analyzing the data, and writing the paper.

Step 1: Recruiting

Whom do you want to participate in your experiment? MTurk offers a heterogeneous participant pool, and requesters can select participants using a qualifications filter. *Qualifications* include, but are not restricted to, geographical location, a worker's HIT approval rate, and his or her number of approved HITs.

- You can combine several qualifications and also negate them. In a more recent experiment, for example, we wanted workers neither from the United States nor from India.

- Be aware that your request of specific qualifications is visible to workers. Consider whether this could influence elicited behavior by inducing private hypotheses among participants about the purpose of your study. You can hide your HIT from workers who are not qualified, but qualified workers may still see your requirements.

- MTurk's *system* and *master qualifications* are meant to ensure data quality mainly for traditional non-research HITs. Be aware that such requirements restrict your sample to experienced participants.

- If the available qualifications are insufficient to filter the sample you need (e.g., parents of children in preschool age), you can first post a HIT that contains only a short questionnaire asking for the relevant characteristics. Such small HITs typically pay only a few cents. You then assign a *self-created qualification* to those who meet your requirements (only visible to you and the qualified worker) and post the actual HIT using the last type of qualification as a participation requirement.

In our project, we used self-created qualifications to prevent workers from participating repeatedly. To restrict participation to workers from the United States and India, we used the existing location qualification. To further attain a balanced sample, we had to put bounds

on the participation rate of American workers who make up the largest share of MTurk's labor force. For each participant pool, we therefore posted two separate HITs per day, one in the early morning and one in the late afternoon (local time). We kept participation for Indian workers unrestricted and, twice a day, only recruited as many Americans as we had recruited Indians earlier that day. This procedure also balances local participation times to equal sized morning and afternoon sessions.

Step 2: Task Description and Informed Consent

Research ethics demand informing participants about a study in advance. But how much is to be revealed? Too much information prior to entering the experiment could strengthen self-selection into participation and influence behavior in the study.

After browsing a list of available HITs, qualified workers can access a *HIT description* prior to accepting a task. This preview serves to attract workers and should allow the check-up of technical requirements. HIT descriptions can also be used as an informed consent sheet. Typically, you provide the following information:

- In our project, we described our *type of task* as "decision experiment."
- For each HIT, you define a *time allotted*, meaning the maximum time a worker has to fulfill the task. Workers not entering their results in time will not be paid. We recommend a generous deadline.
- In addition, you should provide a realistic estimate of the *time needed* to complete the study. Ideally, you establish this time from a pretest.
- A HIT's *reward* is a fixed amount of money for every worker satisfactorily completing the task. In experiments, this amount represents the show-up fee and should attract workers' attention. In our own research, we used show-up fees of US$0.50 or US$1.
- To incentivize decision tasks, you can pay out additional money as a *bonus*. In experimental games, the bonus typically depends on the respective participant's behavior and the choices made by his or her interaction partner(s). To avoid deception of participants, you should be clear about how you calculate the bonus. As stake levels differed in our project, our HIT description just mentioned the chance to earn a "generous bonus" that will depend on the worker's and other workers' decisions.
- *Technical requirements* for participation should display specific software needs (e.g., JavaScript), compatibility with mobile devices, and information on how to restart participation if one accidentally closes the browser window or tab.
- MTurk further allows workers to send messages to their requester through a *contact link*. We included a contact link (connected to our email inbox) to allow for participant feedback both while working on our tasks and after completing the experiment. This dialogue function serves as an evaluation tool in the early phase of your field work. It also helps to decide on whether to approve or reject incomplete work. Particularly

in international online experiments, incomplete submissions may stem from technical problems (e.g., local web connection, software compatibility) and uninformed rejections must be avoided.

- We hosted our online experiment on an external HTTPS-connected website and provided a *link* to this website in the HIT description. To prevent misuse, it is advisable to display the redirection link only after a worker accepted your HIT.

- Finally, we provided an input field for the *completion code* participants received after completing the experiment. For this purpose, we instructed participants to leave open the browser tab displaying the HIT description throughout the experiment. The actual experiment started in a new tab.

Step 3: Worker Information

If your experiment runs on your own website, you may want to transmit available worker characteristics from MTurk. You can include some information as parameters in a personalized redirection link to your own website. Each arrival then carries this information in his or her unique access link, which you can directly store in your dataset. This workaround creates an easy-to-handle interface between MTurk's online platform and your own web infrastructure. Transferable background information includes the following:

- *Worker ID.* This anonymous identifier will tell you who participated in your study and, after completion, will allow individual payment via Amazon.

- *Hard- and software.* Information on a worker's device, operating system, browser type, screen resolution, or JavaScript executability are available from the worker's browser. It allows for adapting the website display to worker's technical needs. We designed our own website to be compatible with different devices, operating systems, and browsers, but we blocked browsers without JavaScript.

- *Qualifications* for HIT acceptance. To validate this particular worker information, it may be useful to recollect them in a questionnaire.

In case you need further worker information for treatment assignment or filtering in the experiment, you may ask relevant questions at the beginning of the study. This information can be used in the process of the experiment. MTurk allows "off-stage" testing of your web infrastructure in the so-called *Requester Sandbox*.

Step 4: Experimental Design

As stated in the beginning, randomization into treatment and control groups as well as the manipulation of treatments are the core elements of experimental designs. In our experiment, we used various levels of monetary stakes as randomized treatments. Our alternative independent variables—participants' country of residence and their regional context—defy

randomization. Those rest on characteristics people bring into the decision situation and thus enter the design only as quasi-experimental "treatments."

Simple questionnaires can run on the MTurk platform itself, which supports a number of question formats in HTML code, CSS stylesheets, and JavaScript. To gain greater flexibility and control over the experiment and its layout, however, it is advisable to host the experiment on an external website, particularly when it comes to randomizing participants into treatments after HIT acceptance.

We further made the experience that building your own online environment from scratch is unnecessary. Instead, various web tools can assist you with the technical implementation of your design. Some, such as *SoSci Survey* and *LimeSurvey*, were developed for web surveys but can also be used for non-interactive experiments. Others, such as *oTree* and *breadboard*, were developed for interactive experiments. Most of them require no or only basic coding skills (in HTML, Java, JavaScript, PHP, or Python), depending on the level of complexity and adjustment needed for your experiment.

In our project, we used *SoSci Survey*, which is free of charge for scientific purposes. One useful feature is "urns" for randomization. For each participant, we drew a specific treatment combination from such an urn without replacement. To deal with dropouts, the urn empties only after a worker's successful participation. When the urn is empty, it is automatically refilled. In combination, this leads to almost perfectly balanced samples, overcoming the problem of misbalances caused by conventional random number generators in small samples. If you change your mind about the number of cases in each treatment condition during data collection, you can refill the urn with alternative entries while still in the field.

Step 5: Instructions

Crowdsourced online research has been criticized because of workers' inattention to detailed HIT descriptions and instructions. The lack of attention is a major challenge for research on MTurk, as its workers reportedly pay less attention to experimental instructions than subjects do in traditional laboratory experiments (Goodman, Cryder, & Cheema, 2012). To mitigate distraction and improve statistical power, we recommend

- *Easy-to-understand language.* English is the commercial language on MTurk, which, in international studies, makes often-problematic translations obsolete. Try to make instructions as clear and easy to read as possible. In our own research, we refrained from multi-language instructions and used simple English, as it is common to almost all MTurk HITs.

- *Illustrations and animations* help to clarify the decision situation and map participants' choices to payoffs as clearly as possible. In our project, we presented animations in GIF format to keep technical requirements low. Because some workers may use a slow or instable Internet connection, illustrations and animations should be designed to cause as little loading time as possible. Also, examples should avoid suggestions of specific strategies or frames to elicit unbiased behavior.

- *Repetitions.* Although short instructions surely help to sustain attention, do not hesitate to repeat essential information you already spelled out in the HIT description. You cannot be sure that opening descriptions are read and understood properly. Many workers shortcut the preview and instantly accept the HIT. We repeated the most important bits of information (e.g., role played and payoff consequences) precisely when participants took their decisions and offered to display the full instructions again.

- *Attention checks.* Typically, you can measure participant attentiveness and understanding with control questions asking for outcomes in various constellations of the decision situation. We placed control questions at the end of the experiment to elicit behavioral data again uninfluenced by specific strategies or frames. The variables generated from such tests permit the statistical control of attentiveness and understanding as well as the exclusion of "suspicious" participants in additional robustness analyses. Moreover, mentioning the requirement to correctly answer control questions to receive payment may help raise attentiveness.

Step 6: Incentives

When thinking about worker compensation, you should consider methodological, ethical, and strategic arguments:

From a methodological standpoint, incentivizing decision-making is an important feature of behavioral experiments. Rewarding selfishness, for example, imposes costs on prosocial behavior. Consequently, behaving fair, altruistic, or cooperative becomes a more valid indicator of prosocial preferences. Of course, this should hold only if relevant amounts of money are at stake. But what is a relevant stake? In our experiment, we randomly assigned each incoming worker to a stake of US$0, US$1, US$4, or US$10. In the US$0 treatment, stakes were just play money and decisions were purely hypothetical, as they did not influence workers' payment. In this treatment condition, workers received a flat US$0.5 bonus, irrespective of their decisions in the experiment. For comparable stakes across countries, we adjusted monetary incentives to differences in purchasing power. Weighted by purchasing power parity (PPP), US$0.4 in India buy as much goods and services as US$1 in the United States. Therefore, workers from India received stakes of either US$0.4, US$1.6, or US$4, which provide the same purchasing power as US$1, US$4, and US$10 in the United States. Empirically, it turned out that it is important for Dictator Game behavior to provide monetary incentives. Workers in the US$0 treatment showed unrealistically high rates of altruism, and rates decreased once we introduced real monetary stakes. The actual size of the incentive, however, has no effect on behavior. In line with a pioneering study by Amir, Rand, and Gal (2012), we thus concluded that small stakes suffice to induce both attentiveness and self-interested behavior at MTurk.

Vanessa Williamson (2016), a political scientist at the Brookings Institution, argues that workers on MTurk do not participate in scientific studies just for fun but that many depend on additional income to meet their basic needs. Researchers should thus pay at least the U.S. minimum wage of US$7.25 an hour. In our experiment, we paid a show-up fee of US$0.5 to each worker completing the experiment plus the money earned in the experimental games.

U.S. workers on average earned US$2.34 in 14 min. This is a better payment than most requesters at MTurk offer. Still, some workers in our experiment received only US$0.5. At least, every worker had the same chance to get into a high-paying treatment condition.

As a researcher, you want to attract workers for your experiment and, at the same time, save research funds. Consider that workers screening the list of available HITs (before they can access a more specific HIT description) only see the *reward* (your show-up fee) but not the *bonus* (your stakes) and only the *time allotted* (the deadline) but not the duration of the experiment. Many workers are active on Internet forums such as *Turkopticon*, *Turker Nation*, and *MTurk Forum* to exchange information about HITs and requesters. If you pay workers well, word will spread about your HITs. On the other hand, the most challenging problem with respect to forums is crosstalk. If in our experiment, for example, participants exchanged information about monetary incentives, they could have found out quickly about our stake variations and thus influence the behavior of future participants. We recommend keeping one eye on the forums, treating workers with respect, and shortening field time for your experiment.

Step 7: Confounders

Due to randomization into control and treatment groups, experimental results should not be biased by confounding variables. Yet, to generalize your results to larger groups of people, you need to statistically control for potential differences between your sample and the target population. Such adjustments in your data analysis are similarly important if you compare the behavior of different demographic groups within your experiment, particularly, if self-selection into the experiment works differently across groups. Remember that such comparisons rest on non-experimental variation and remain mostly descriptive, even when controlling for all known confounders.

MTurk workers are routinized and willing to provide information on their sociodemographics. At the same time, many workers are experienced study participants, and familiarity with social science theories and designs might jeopardize internal validity. In our project, we asked workers to report the number of experiments they previously participated in. We also collected information on individual motivations to participate (e.g., money, scientific interest) as well as the surroundings during participation (e.g., location of participation, being watched by others).

When you combine an experiment with a survey, we recommend administering the questionnaire at the end of the experiment. A post-decision questionnaire cannot influence elicited behavior, and participants have no incentive to misreport to increase their (assumed) chances of participation. If some survey data seem implausible (e.g., cases with very high income), consider dropping or conditioning on those cases in additional robustness analyses. When all this is taken care of, your data will permit careful comparison of different groups and may unveil interesting results impossible to find in traditional laboratory experiments.

Step 8: Matching and Payment

For simple dyadic situations in which only two participants react to each other, it is more comfortable for the workers and more feasible for the experimenter to break up synchronous interaction. The experimental games we used permit such a desynchronization of participant interaction. In the DG, half of our participants were passive receivers anyway. In the UG, we had each proposer take his or her decision while each responder autonomously stated acceptance or rejection of each feasible offer (0%, 10%, ... , 100% of the pie). Using this "strategy method" (Rauhut & Winter, 2010), one can calculate a responder's minimal acceptable offer (MAO) considering all possible proposer decisions. We stayed away from the standard measure of responder behavior (second player's response to proposer's actual offer) also because offers often exceed responders' MAO (censoring the response variable).

Refraining from real-time interaction, each worker participated separately. We only paired participants to calculate payoffs after they completed the HIT. Ex-post matching avoids waiting time and dropout from the experiment. Under positive stakes, payoff included the show-up fee (US$0.5) and a variable bonus computed from the focal participant's (and partner's) decisions. To do so, we matched each finalist to a complementary decision randomly drawn from the pool of preceding participants (without replacement). We then stored any remaining unmatched decisions for future matchings.

With this workaround, workers receive payoff feedback right after participation. We provided an anonymous completion code on the same screen, which workers then used to collect payment through their Amazon accounts. In your own project, you should store the completion code in your dataset together with the worker ID and verify that both the ID and the code match the identifiers submitted at MTurk. You may then send the *reward* by *approving* the work and transfer the *bonus*.

After participation, workers expect payment as soon as possible. Be cautious about rejecting a worker's submission. We chose to first get in contact with problematic submitters by email (you can find workers' contact details in the rubric "Manage HITs individually" under "Download results" by clicking on the worker ID). A rejection does not only mean not being paid but may also ruin a worker's status: Non-completed or disapproved HITs may limit a worker's qualification for future HITs. Many requesters use a 99% HIT approval rate as an entry qualification, and workers need 100 approvals to override a single rejection. If you want to send money to a worker who failed to submit, use a *dummy HIT* and a corresponding *self-created qualification*.

Step 9: Preventing Repeated Participation

Making sure that each worker can participate only once is an important aspect of data quality. You can prevent repeated participation in several ways:

- The simplest way, implemented in MTurk, uses the *worker ID*. Each worker can, by default, accept each HIT only once. However, if you repost the HIT, the same workers can accept it again.

- Amazon further suggests rewarding each participant a *self-created qualification*. Only workers without this qualification may participate in later HITs. Gabriele Paolacci (2015 in our web resources) provides a useful Excel tool for partly automating qualification assignment. A worker's qualification is visible only to you and the worker.

- With more than one similar HIT running synchronously, qualifications alone may not suffice. In addition, you can use a *database* to keep track of all worker IDs entering your experimental website. Worker IDs are collected from each arrival's personalized redirection link and workers with known IDs are blocked.

Because many workers will forget which HITs they had previously worked on, we built a small website where workers could enter their IDs and check whether they are eligible for participation. We provided a link to this online tool in our HIT description. This is important, as workers otherwise would only learn about blocking after HIT acceptance, which could hurt their HIT approval rate.

Step 10: Accounting

University administrations' accounting procedures are often well adapted to laboratory experiments. There, each participant signs a receipt upon receiving payoff, which the experimenter hands in together with a list of overall expenses. This procedure is less suitable for the accounting of crowdsourced online experiments. At MTurk, participants can be identified only by their worker ID.

To start experimenting at MTurk, you first need a requester account registered with your credit card number. Second, you must make a prepayment for your HIT. On the requester website under "My Account: Transaction History" you will find

- prepayments for HITs with your credit card,

- payments to workers,

- payments of Amazon fees.

If you zoom in on single field days, you will also find the amounts sent out to each worker. In our case, and due to the kindness of our local university administrators, we could hand in screenshots to prove which worker received how much money. We then verified total payments with our credit card bill. If you only field one project at a time, this should be sufficient for your accounting of research expenses. You should sort out accounting details with the cashier's office at your university or your funding institution before fielding your experiment.

Final Remarks

Start into the field with caution. When you think your instrument for data collection is finished, pretest it first. Ask colleagues and students to try it and note obscurities. Also check the understanding in cognitive pretests, for example, by requesting early participants

to paraphrase instructions. Use the Requester Sandbox to test your infrastructure on MTurk. When everything seems ready, start with only a few workers. During such a "soft launch," check for unwanted behavior of the website, systematic dropouts, and participants' comments before having more workers participate.

Evaluate data quality. One indicator of data quality is answering time. Very fast participants may not have considered the task thoroughly. Very slow participants may have looked up eligible answers and best strategies. Instead of arbitrarily excluding outliers, mark these cases to statistical control for in your data analyses. Other indicators of low data quality include inconsistent behavioral choices, many repeated trials to answer control questions, implausible answers in the questionnaire, postings of pointless comments, and temporary absence from the website.

Get help. Below we have listed important web resources that are helpful for your own web implementation. Also note that many experimenters supplement their papers with technical appendices and experimental instructions. To implement your own design, these supplements are often even more valuable than the paper itself. If you need further assistance, we can provide code for MTurk HITs, SoSci Survey questionnaires, randomization, instructions, prevention of repeated participation, and ex-post matching upon request.

We hope this report is a helpful starting point to find your own way of unveiling the potential of crowdsourced online experimentation.

Acknowledgement

This project has been supported by generous grants from the German Research Foundation (KE 2020/2-1).

Exercises and Discussion Questions

1. List all preconditions needed to technically implement an online experiment at Amazon Mechanical Turk.

2. Find your own research question that can be examined in a crowdsourced online experiment. Why is an online experiment more suitable for this research question than other methods of data collection?

3. Go through the 10 steps with your research question from exercise 2. Which design decisions need to be taken at each step? How do you decide and why so?

4. Think about the disadvantages of online experiments. Are you still convinced that the advantages outweigh the disadvantages?

Further Reading

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

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46, 112-130.

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44, 1-23.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23, 184-188.

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

Web Resources

Chen, D., Schonger, M., & Wickens, C. (2016). *oTree: An open-source platform for laboratory, online and field experiments*. Available from http://www.otree.org

Experimental Turk (2016, October 18). A blog on social science experiments on Amazon Mechanical Turk. *Experimental Turk*. Retrieved from https://experimentalturk.wordpress.com

Ipeirotis, P. (2010). Analyzing the Amazon Mechanical Turk marketplace. XRDS: Crossroads. *The ACM Magazine for Students, 17*(2), 16-21. Retrieved from http://demographics.mturk-tracker.com/#/countries/all

McKnight, M., & Christakis, N. (2016, May 1). *Breadboard: Software for online social experiments* (Version 2: Yale University). *Breadboard*. Retrieved from http://breadboard.yale.edu

MTurk (2017). *FAQs*. Retrieved from https://requester.mturk.com/help/faq

SoSci Survey. (2017). *Create online questionnaires and run surveys on the Internet*. Retrieved from https://www.soscisurvey.de/index.php?l=eng

References

Amir, O., Rand, D., & Gal, Y. (2012). Economic games on the internet: The effect of $1 stakes. *PLoS ONE*, 7, e31461.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonna'iveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46, 112-130.

Goodman, J., Cryder, C., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26, 213-224.

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

Keuschnigg, M., Bader, F., & Bracher, J. (2016). Using crowdsourced online experiments to study context-dependency of behavior. *Social Science Research*, 59, 68-82.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, 5, 3677.

Rauhut, H., & Winter, F. (2010). A sociological perspective measuring social norms by means of strategy method experiments. *Social Science Research*, 39, 1181-1194.

Williamson, V. (2016). On the ethics of crowdsourced research. *Political Science & Politics*, 49, 77-81.