A User's Guide to Collecting Data Online

Kalif E. Vaughn

Jeremy Cone

Nate Kornell

Author Notes

Kalif E. Vaughn, Department of Psychological Sciences, Northern Kentucky University

Nate Kornell, Department of Psychology, Williams College

Jeremy Cone, Department of Psychology, Williams College

Abstract

Although psychological research has been traditionally conducted in laboratory settings, online data collection has become increasingly popular in recent years. Online data collection offers a variety of benefits over laboratory-based data collection, including the ability to recruit large sample sizes, improve participant accessibility to research, and collect data at breakneck speeds. Despite these advantages, researchers may have concerns about the quality of data collected. Furthermore, researchers may be interested in collecting data online but have a limited understanding of the process. We discuss research that suggests the quality of online data is equal to that of laboratory-based data, particularly if one follows the various safeguards that we address. Additionally, we provide an overview of how to get started conducting research online, including practical tips such as how to create your experiment, how to advertise your experiment, how much to pay your participants, and how to word your instructions. Although these tips apply to all online data collection, our examples and remarks center around Amazon's Mechanical Turk service.

Why collect data online?

The internet allows researchers to collect experimental data without having participants come into the lab. Collecting data online has numerous advantages compared to traditional lab-based data collection. For example, having one hundred participants complete an experiment in the lab might take several weeks or even an entire semester depending on the resources available. If run online, the same experiment typically takes less than a day and requires no additional labor beyond setting up the experiment. Furthermore, to double the sample size (i.e., run 200 participants) in the lab would take twice as long and much more work on the part of the research staff, but doubling the sample size online takes no more effort to collect the data and hardly any more time.

Perhaps surprisingly, online data collection is not overly expensive. As we will outline later, the typical experiment may only cost a few hundred dollars to run. Obviously, it depends on the length of the experiment and how many participants you need, but achieving large sample sizes for a modest cost is entirely feasible with online data collection. Furthermore, most cognitive and social cognition studies, including memory experiments, can be conducted online. For example, experiments wherein participants learn information (e.g., word lists, word pairs, text materials, etc.) and are subsequently tested can easily be conducted online. Additionally, collecting survey information (e.g., gathering public opinion about a topic, asking people to make predictions, etc.) is also well-suited for online data collection. As we will outline later, even complex experiments can be run online, and clever paradigms have been created to ensure that participants complete the tasks correctly. (We also point out later that some standard paradigms and scales need to be used carefully because so many mTurk workers have encountered them before.)

Who is this chapter for?

If you have no experience collecting online data, this chapter is for you. We provide systematic instructions on how to run a study online. For readers who have collected data online before, we offer suggestions about best practices that might be useful to consider. Our goal is to make this chapter useful regardless of your skill with technology.

Technical setup

People sometimes think of "Mechanical Turk study" being synonymous with "online study." It is not. Online studies involve two basic components: recruiting participants (often but not always on Mechanical Turk) and collecting data (almost never on Mechanical Turk®).

Recruiting Online Participants

Amazon's Mechanical Turk® (mTurk) allows researchers to pay workers for completing "Human Intelligence Tasks" (HITs). There are other "crowdsourcing" alternatives such as Qualtrics® and SurveyMonkey®, or participants can be recruited for online studies through the subject pool at one's college or university.

Choosing an Online Data Collection System

Although data can be collected directly on mTurk, it is rare and only happens with simple studies. For complex studies, after a participant has decided to do your HIT, they typically click a link, provided by you, that sends them to a separate website. This separate website hosts your experiment and is where the data are collected.

There are various approaches to data collection. Which method is best for you? It depends on what you need to accomplish and your level of technological expertise. Simpler systems are less flexible and allow you to do less, but they are easier to use and sometimes less error-prone. At the simplest extreme, Google Forms and SurveyMonkey are user friendly. They are also limited in what they can do. If you want to use random assignment, collect reaction times, or shuffle the order in which questions are asked, you will need to use a more complex platform. For most researchers, especially those trying online data collection for the first time, the best approach lies in the middle. For example, Qualtrics is much more powerful than a Google Form, but it is user friendly and does not require writing code. There are also collaborative systems with open-source software such as PsiTurk, which allows for even more customization but requires coding knowledge. At the more technically demanding extreme, you can set up a website of your own by registering a domain name (e.g., www.yourdomainname.com) and setting up a web server (usually at a small yearly cost; as of this writing in the $200 range). You can then write code on the server (e.g., using PHP, JavaScript, or HTML5) that will run whatever study you want.

If you choose to write your own code, there are frameworks (basically, downloadable code) you can use as a starting point, such as a system called Collector. (The third author is a contributor to the Collector code.) Such systems take a set of instructions (written in spreadsheet form) and turn them into a program for you. For simple projects, these frameworks require little or no coding, but you can modify the code to suit your needs. Some experiment presentation software packages also offer online study options, including Inquisit. Such pre-packaged software usually incorporates many of the procedures that are necessary to execute an experiment, such as accurately recording reaction times and writing them to file. Because it has been thoroughly debugged and tested, it is also less prone to errors that could invalidate the conclusions of your study.

A final consideration is the company or group that controls the system you choose. Large companies such as Google and Qualtrics are stable, provide good documentation and examples, and offer support and example experiments in online forums. Moreover, they are not likely to go out of business, or stop supporting their code. A smaller outlet (such as Collector, which we mentioned above) has less documentation, less financial backing, and thus, may have a greater chance of disappearing. We note, however, that Collector can provide an excellent segue into learning how to code in HTML5, PHP, and JavaScript, which are languages likely to be supported within web browsers from now into the distant future. PsiTurk is an appealing option because it is a collaboration between many different labs, each of which can add to the codebase, and it also appears to have reached a point where the platform is stable and widely used. Our recommendation is to use the simplest solution that will work for you. Trying to use code that you do not understand leads to mistakes, and even the savviest programmers are sometimes better off with a simple Google Form (links to these services are listed in Appendix A).

What to say on your recruitment page

Regardless of whether you are recruiting participants on mTurk, a subject pool website, or elsewhere, there are guidelines to follow.

Provide a brief description of the experiment

It is not necessary to go into detail, as long as the participant knows what he or she is (and is not) signing up for. The detailed instructions will be presented later (including inclusion/exclusion criteria and a full description of the experiment). For example, we have used descriptions as simple as “answer trivia questions” or “learn word pairs and then take a test on them.” Of course, it is prudent to follow any guidelines and recommendations set forth by your institution’s IRB.

Indicate roughly how long the experiment will take

We sometimes do this by specifying on the recruitment page how many minutes it would take to complete a study. For a self-paced study in which participants answer 100 trivia questions, our recruitment page might say the experiment will take between 15-30 minutes. Using an interval is desirable because participants will vary in how long they take to finish the study. In our own work, we generally ask a research assistant who is unfamiliar with the experiment to time themselves as they complete the experiment to get a relatively accurate measure of the length of the study.

Tell participants how much compensation they will receive

In every online recruitment system that we know of, you have to indicate compensation in order to recruit participants. Mechanical Turk includes this information in the advertisement for the HIT, but it can be helpful to reiterate this information so that it is readily available to workers.

When running a two-session experiment, we usually do not mention the second session when recruiting participants for the first session (although we do mention it when using a subject pool), because there is a high likelihood that some participants are not going to be asked to complete session 2 (e.g., if they did not follow the directions). Thus, we only promise to compensate each participant for the first session (which is mandatory). When the experiment is over, we tell them that we might invite them back for a second session. Given that online experiments ease the burden associated with collecting large amounts of data, it is possible to exclude a fair number of participants upfront (e.g., for noncompliance) and still have ample data for your experimental needs.

Data points you should collect for administrative reasons

There are two pieces of information you will need to know about the participants that you recruit: a) who they are and, b) whether they successfully completed the experiment you are running (see below).

Who your participants are

Assuming your data are collected separately from your recruitment system, you will end up with two separate sets of data. You need to know which participant in the recruitment data is which participant in the actual data file.

For example, assume you have recruited participants using mTurk and run a study on Qualtrics. A given Qualtrics participant might be called “participant 1” but on mTurk they might be called “A1T834UD.” Why do you need to know that A1T834UD is the same person as participant 1? There are multiple reasons. First, you may want to compensate participant 1 but not participant 2 on mTurk. Second, you might want to offer participant 1 additional compensation for their work in your experiment (i.e., give a “bonus” in mTurk), which requires you to know their mTurk ID. Third, if you wish to exclude participants who have participated in previous studies, the only means of tracking their participation in the system is through their mTurk ID.

We suggest that you ask the participant to enter their mTurk ID as soon as they get to Qualtrics, or whatever system you are using to collect data. This way their mTurk user ID will be the same in both of your data files. Mechanical Turk workers regularly enter this information when completing HITs and therefore expect it to be requested of them.

Did the participant complete your whole study?

On the last page of a study, researchers typically give participants a “completion code,” which is usually a random collection of letters and numbers that workers cannot know without having completed the study. Participants should then be asked to enter this code on the mTurk page before they “submit” the HIT. (Note that mTurk will allow participants to reserve a spot for themselves in the HIT--or “accept” the HIT--and then complete the study before they enter their completion code and submit the HIT.) mTurk participants expect to see a completion code at the end of a study (and send us concerned emails if we do not include one).

Note that this information can be critical not just for providing compensation but also for theoretical reasons. Some recent work (Zhou & Fishbach, 2016) suggests that it is important to attend to attrition in online experiments in order to ensure that the conclusions of your online experiment are valid (e.g., that participants who finish the experiment are not qualitatively different than those who drop out of the experiment).

To summarize a typical sequence of events using mTurk, participants “accept” your HIT, click your link, enter their mTurk ID on the first page of your study, complete the rest of your study, get a completion code at the end of your study, go back to mTurk, enter the completion code, and then submit the HIT. (Note: For the convenience of your mTurk workers, it is best to have the link on mTurk open in a new tab or window, so that when participants finish your study the mTurk tab they started on is still open.)

Writing instructions

Writing clear instructions is of paramount importance when running any experiment, and online experiments are no exception. The instructions should communicate precisely what the participant is supposed to be doing throughout all phases of the experiment. Below, we offer guidelines to improve the quality of your instructions.

Be concise

Do not make your instructions look like the iTunes terms of service (i.e., overly long and boring). Most internet users have a habit of moving through web pages quickly and skipping over long paragraphs. Users are most likely to read instructions that are brief and direct.

Give an overview at the start

Give participants a clear outline of the procedures that they will be completing during the study. Make it as specific as possible. Be sure to include how long the experiment will last, as well as explain the various tasks they will be completing and what you would like them to do during those tasks.

Provide instructions before any new sub-task

At the start of each new phase of your experiment, it is beneficial to provide a brief set of instructions that reminds participants exactly what the next phase of the experiment entails. By providing these instructions, you will make your participants feel that the experiment has more continuity and is easier to complete. Moreover, you should never underestimate their power to forget or misunderstand your initial instructions.

Be repetitive

It is important to err on the side of being repetitive. For instance, you may have already defined a particular term or outlined a particular procedure several times already, but it is worth repeating it rather than assuming that your participants will remember on their own.

Administer comprehension checks

At any phase of your experiment, you can administer comprehension checks. For instance, to check whether your participants understood your instructions, have them type out a summary of the instructions they just read, or give them a multiple-choice test (e.g., “What should you be doing during this phase of the experiment? Check all that apply.”) This allows you to keep track of which participants understood the instructions, and can also help identify any common misunderstandings (such that you can use more effective prose later on).

Be careful about excluding participants based on whether they followed instructions as this can sometimes lead to a biased sample(see Tinghog et al., 2013, and the associated reply from Rand et al., 2013, for a discussion of this issue on manipulations of response time. Also, see the section on *Excluding data* for more on this topic.)

Give reminders on each trial

Put a brief note that explains what your participant should be doing on each specific trial (e.g., "Recall and type in the word in the box below"). This is an additional safeguard in case your participants misunderstood or forgot the initial instructions.

Avoid jargon

Write your instructions in plain and simple English. If a specialized term must be included, then be sure to explain what it means using clear language and by providing at least one example.

Quality assurance

There are a variety of ways to improve the quality of the data obtained from mTurk, which we discuss below.

Clear descriptions

Quite often, when an mTurk workers fails to complete the task in a satisfactory manner, he or she has simply misunderstood it. In our collective experience, we rarely encounter participants who are outright negligent and trying to game the system, in part because they will lose work if they get a bad reputation. Therefore, you will want to write instructions that are clear and unambiguous (the prior section can help you with this).

Filter out mTurk workers who do not have a strong reputation

Within mTurk, you can restrict participation in your study using a variety of filters. Two of the most popular filters are completion rate (i.e., the percentage of HITs for which the worker has been approved) and the number of HITs completed by the *mTurk workers*. We only allow *mTurk workers* to complete our HITs if they have at least a 95% approval rate. We sometimes require that they have completed 1000 HITs.

Filter out mTurk workers living outside the United States

You can filter out mTurk workers based on the country where they live. This setting can help avoid mTurk workers whose first language is not English or who are not fluent in English). Unfortunately, there are methods to obscure geographical location on the internet, such that workers can make it appear that they are located within the United States. If using native speakers of English is an important consideration for your research, it may be necessary to use this filter in conjunction with manipulation checks about the English language (e.g., difficult rhyming tasks or questions about common slang terms in English).

Pay mTurk workers a reasonable sum

As explained in more detail in *How much to pay,* paying mTurk workers fairly will help you recruit quality participants and maintain a good reputation.

Avoid long and boring tasks

mTurk workers are ordinary human beings who are subject to fatigue and human error. If you want participants to stay focused on your task, you should consider keeping it short and to the point. At a maximum, we do not recommend creating tasks longer than 1 hour unless it is an interesting and engaging task. Typically, our studies range from 10-30 minutes. We recommend having RAs complete the study (or piloting your study on mTurk) and analyzing how long it takes for most participants to complete the task. This will not only help you determine the length of the experiment, but it will also give you an idea of how much to pay your workers.

Requiring a minimum trial time

Some of your participants will be tempted to take less time per trial than they should in order to finish your experiment quickly. This happens in the lab, but it might be more common online (perhaps because participants are not being monitored in person). We sometimes impose a minimum trial time to insure that participants take the time to process our questions. For example, we might show participants a question and a box where they can type their answer, at the start of a trial, wait six seconds, and then show them a button that they can use to submit their answer and end the trial.

Identify negligent mTurk workers using comprehension checks

You might want to use random comprehension checks to identify *mTurk workers* who are not paying full attention to your task (Kittur, Chi, & Suh, 2008). Inserting questions designed for this purpose is desirable for two reasons. First, it will allow you to exclude participants who were not focusing on the task. Second, you will be able to define exclusion rules objectively prior to running the study (e.g., they should be excluded if they get questions A, B, or C wrong). This is preferable to a situation where you end up having to use subjective opinion to exclude participants (see Simmons, Nelson & Simonsohn, 2011).

One way to create such checks is to ask questions that are relatively simple, but only if the participant has been paying attention. For example, Downs, Holbrook, Sheng, and Cranor (2010) posted an email message, which was disguised as a formality of informing participants about the qualifications needed to participate in the study. MTurk workers were asked to answer an easy question and a difficult question about this email message. The easy question could be answered by simply looking up the answer in the email, whereas the difficult question could be answered only after a careful reading of the message. In other words, an incorrect response to the difficult question suggested that the worker only skimmed the message.

We recommend writing a unique set of comprehension checks for each study you run, because some experienced mTurk workers can spot familiar comprehension check questions. One mTurker said “Whenever I see the word vacuum, I know it’s an attention check,” (Marder & Fritz, 2015).

How much to pay

Although there is a wide variety of pay rates within Amazon’s mTurk, we recommend erring on the side of overpaying participants rather than underpaying them. Paying mTurk workers fairly is valuable in its own right. It will also help you maintain a good reputation (see *Maintaining a positive reputation*), and it helps to establish trust between the researcher and the worker. Not only are these mTurk workers more likely to take the task seriously, they are more likely to participate in your future HITs.

One of the best ways to ensure adequate compensation is to thoroughly understand your experimental design, and in particular how much time and effort your experiment will require from the participants. Piloting your study on naïve research assistants will give you an accurate estimate of how much time the experiment requires. If you determine the experiment takes around 22 minutes during the pilot process, it is a good practice to list the experiment as taking a little bit longer than that when advertising it online (e.g., 25-30 minutes) and pay accordingly. This ensures that the slower participants are still compensated adequately. *Note*: Currently, mTurk will tell the worker the “effective hourly rate” for HITs that he or she has completed. This number is simply the amount you pay divided by the amount of time Amazon thinks your participants spent doing your study. This metric is frequently inaccurate because certain participants do not submit the study immediately after finishing it, or do not start as soon as they accept it and so forth. Thus, the “effective hourly rate” provided by Amazon is often less than the workers’ actual hourly rate (at least in our experience).

When and how to pay workers

Compensating workers, in mTurk language, is the same as approving workers for a specific HIT. Each HIT will have a special section that lists all the users who have completed the HIT and are awaiting compensation. You can pay workers manually or automatically.

Paying mTukers manually

Paying workers one by one can be desirable because it gives the experimenter time to verify each username against his or her data. If the worker did not successfully complete the HIT, you could avoid paying the worker by rejecting him or her (although we do not recommend it; see below in “Maintaining a positive reputation”).

Paying mTurk workers automatically

Although you can manually pay workers for completing your HIT, it is more efficient to pay the workers automatically. mTurk can automatically approve workers after a fixed amount of time (e.g., 5 minutes). Automatically approving workers saves time and, perhaps more important, mTurk workers appreciate being paid quickly. If the mTurk workers enjoy your HIT and are paid quickly, they are more likely to participate again in the future and they are more likely to provide positive feedback about you as an experimenter (e.g., on an online forum). The drawback is that you will end up paying all participants, even those who did not complete the task correctly. In our experience, the benefit of doing this (a better reputation, less labor, and fewer complaints) outweighs the cost (paying a few dollars to workers who do not deserve it), but it depends on whether the experiment is for pay or credit and how much the pay is.

Giving mTurk workers bonuses

Giving bonuses is useful for multi-session studies. We typically pay workers for session 1 and then use bonuses to pay them for additional sessions. Bonuses can also be used to give workers incentives for doing well in the task or for other reasons. As we explain in the section on *Multi-session studies*, mTurk workers like bonuses a lot. In one study, more signed up to receive a $1 bonus than signed up to do a new HIT worth $2 (Stoycheff, 2016).

Troubleshooting payment

Be prepared to get messages from mTurk workers who completed your HIT but were not paid. Usually this happens because they are unable to submit the HIT, and this could be because (1) they closed the browser window with mTurk in it, (2) they forgot to accept the HIT (which means they did not reserve a place for themselves to do the HIT and found that it was already over when they tried to submit), or (3) they did not submit within the designated time window. Typically, this participant will show up in your data even though they are not visible on mTurk.

When a participant tells us that he or she completed our HIT, our policy is to give them the benefit of the doubt and try to compensate them if we can. This can require some effort, however, because as of this writing, mTurk will not allow you to pay an mTurker unless they have submitted a HIT for you. If they have submitted a HIT for you (perhaps in a different study) you can pay them with a bonus. If not, we sometimes create a HIT for a single participant. We put that participant's mTurk ID in the HIT description. We make sure the HIT does not pay anything because otherwise, other workers will sign up for it. If the right worker submits the HIT, we approve him or her and then give a bonus for the amount we owe. Doing this is time-consuming for the researcher, but it is the right thing to do, and it will help you maintain a positive reputation.

Giving your money to mTurk

Note that all experiments must be prepaid on Amazon: The experimenter is required to have a "balance" on Amazon before any HIT can be created. Once you create the HIT, Amazon pays your workers directly from your pre-existing balance.

Maintaining a positive reputation

If participants like you as an experimenter, they will be more likely to complete your future experiments and to provide high-quality data for you. Furthermore, they may recommend you to other online workers. mTurk workers use third party systems to rate requesters (i.e., people who post HITs) such as Turkopticon (see Appendix A). If you get a bad reputation on such a site, it might become difficult to recruit participants. There are several ways to maintain a positive reputation.

Compensate workers appropriately

There are thousands of other tasks that workers can complete online at any given time. All other aspects being equal, studies that pay more will be more desirable for workers. We strive to pay workers a minimum of $6 per hour on mTurk. When in doubt, err on the side of overpaying workers (if feasible). We have also found that mTurk workers are more sensitive to the amount they will be paid than their pay rate; for example, they will sign up faster for a study advertised as $3 for 30 minutes ($6/hour) than for a study that pays $0.50 for 3 minutes ($10/hour).

Be responsive to email

Workers will often contact you using the email address you provide to sign up for the service if they run into any trouble with the HIT. They value prompt feedback and communication with requesters so that they can resolve issues quickly and finish their work efficiently. In our research, we generally strive to respond to emails within 15 minutes or less to ensure that we always promptly resolve technical issues or study-related questions. Workers have good reason to be wary, because there are predatory requestors on mTurk who pay unfairly and at low rates.

Compensate workers quickly

As we explained in the section on *When and how to pay your workers,* workers are not automatically paid by default, but you can change this setting so that workers are automatically paid. Workers appreciate being automatically paid, and it is an easy way to maintain a good reputation. If you do pay automatically, it is still possible, and desirable, to pay workers promptly.

Make your study interesting

Regardless of what your study is about, there are a few simple ways to make it more interesting. First, avoid long studies. No matter how terrific your study is, workers would probably rather be doing something else. Second, use fun distractor tasks. If you need participants to complete a distractor task for 5 minutes, let them play Tetris instead of doing math problems, unless, of course, the type of distractor task is critical to the study. Third, make your study visually appealing. You can do this by using crisp fonts, high resolution images, appropriately-sized boxes and menus, and so forth.

Only reject workers when necessary

Occasionally, workers may fail to complete your experiment in the manner you had intended. As the experimenter, you have the option to reject their work and deny their payment. However, rejections are recorded in their worker profile and count against their approval rating. Workers with low approval ratings (e.g., below 95%) can have trouble getting future assignments. Therefore, unless a worker blatantly disregarded your instructions and did not exert any effort to complete your experiment (e.g., the participant just “mashed” buttons, that is, responded without attending, throughout the experiment), we recommend approving all assignments and paying the workers. This will save you valuable time because you are not checking each data file individually for noncompliance (at least, not during the approval process), and will ultimately improve your reputation.

Multi-session studies

Running a multi-session study in the laboratory is a tedious and time-consuming task. First, you have to recruit participants manually, either via flyers or perhaps advertising using your institution’s research collection service (e.g., SONA systems). Then, you have to manually run each participant in the lab. Of course, there are physical limits with respect to how many participants you can run at the same time depending on the size of your lab space. For this example, let us assume that you can run 6 participants at one time. If you are aiming to collect 60 participants for your first experiment, and *the first session of your first experiment* takes 1 hour to complete, then the minimum amount of time required to collect the data for the first session will be no less than 10 hours. If subsequent sessions also take 1 hour to complete, and you want to run participants for a total of 5 sessions, then the total time to collect data for your first experiment will be a minimum of 50 hours. To state the obvious, this is not an easy task! Additionally, this is not including recruitment time (e.g., waiting for participants to sign up for the experiment) nor is it factoring in attrition rates, which is a perennial issue in any sort of multi-session experiment (e.g., Gustavson, von Soest, Karevold, & Røysamb, 2012).

Contrast running a multi-session study in the laboratory to running a multi-session study on mTurk. With mTurk, you start by posting your description of the experiment and available HITs for the community to view. Then, participants complete the first session of your experiment on their own, electronically, without any additional effort on your part. Realistically, you could have your entire first session sample within a few hours, depending upon the rate of pay and the difficulty of the task. After your first session is complete, the follow-up sessions are completed rather easily. You simply email the mTurk participants a link to your experiment after the appropriate time has elapsed (e.g., 1 week if that is your desired delay). These emails have to be sent through the mTurk page because you will not know the participants' email addresses. Participants complete the remaining sessions just as they completed the first session, and your experiment is completed with a minimal amount of effort.

In addition to the reduced resources required and the time saved, running multi-session studies on mTurk offers other unique benefits as well. Although attrition is a problem in any experimental context requiring multiple sessions, mTurk has the strong possibility to lower attrition rates compared to laboratory studies. One obvious motivator for improving attrition rates is to offer an increased financial incentive to participants (e.g., pay them additional money for returning to each session). However, one study found evidence that participants do not choose to complete follow-up sessions based solely on financial reward. Stoycheff (2016) had mTurk participants complete a short 5-minute survey and then emailed them to return a week later to complete a second 5-minute survey. Of interest, Stoycheff manipulated the context of the follow-up email. Some participants were offered $0.50 to complete Part 2, others were offered

$1.00 to complete Part 2 but had to complete a novel HIT, and yet others were offered $1.00 to complete Part 2 but were awarded the money as part of a bonus payment on the original HIT. There were no significant differences in the return rate between the $0.50 condition and the $1.00 conditions, suggesting that participants are not necessarily motivated strictly by financial incentives. However, there was a significant increase in the return rate for participants offered a $1.00 reward when it was awarded as a bonus payment compared to a new HIT. Stoycheff posits that even in cyberspace, by agreeing to complete multi-session experiments, participants are establishing a social relationship with the experimenter. Bonus payments may remind participants of their original agreement with the experimenter. mTurk workers might also trust you more because you have already paid them for their work in session one. Furthermore, bonus payments may also have an inherent social value as they are typically awarded for exemplary performance. Therefore, we recommend using bonus payments to further improve retention rates on mTurk.

One oddity that you might encounter is participants who complete session 1 but do not submit your HIT. Because you cannot contact these participants, you will not be able to invite them back for session 2. One other option to consider is running session 1 in the lab but doing the follow-up study online. This would not involve mTurk, but if you need to do a long or complex first session and the subsequent sessions are simple (e.g., a memory test), it can be a good option. In sum, mTurk can reduce the time and energy required to complete multi-session experiments. Furthermore, mTurk can lead to excellent retention rates in multi-session studies, even up to 80 or 90% (e.g., Christenson and Glick, 2013; Bartels, 1999).

When mTurk workers are not naive

There are professional mTurk workers who estimate they have completed over 10,000 academic studies (Marder & Fritz, 2015). Although these workers might be relatively small in number, they may show up disproportionately in your HITs. One study examined over 16,000 mTurk participants across 132 batches of HITs that were used for academic research. The results showed that the average mTurker had completed 2.2 of these HITs, and the most prolific 10% of the mTurk workers had produced 40% of the responses (Chandler, Mueller, & Paolacci, 2014).

If you are making use of a commonly used scale, set of materials, or paradigm, do not assume that your mTurk participants are naive. Some of these mTurk workers have completed standard scales, such as a self-esteem scale, many times. They also know the answers to many of psychology's trick questions, such as "A bat and a ball cost $1.10. The bat costs one dollar more than the ball. How much does the ball cost?" and some are so experienced that they even know that this question is usually followed by a question about a widget and then a question about lily pads (Marder & Fritz, 2015).

A common example in memory research would be materials from the Deese-Roediger-McDermott paradigm, which are composed of lists of words (e.g., *hot*, *snow*, *warm*, *winter*, *ice*) that make people think of a critical lure (e.g., *cold*; Roediger & McDermott, 1995). These lists are used to study false memories, and are popular enough that we recommend exercising caution if using these types of materials online given their popularity. Similar caution is in order when using stimuli from other studies that have become very well known, such as when showing the famous invisible gorilla video (Chabris & Simon, 2010) or asking people "how fast were the cars going when they smashed into each other?" (Loftus & Palmer, 1974).

Experienced mTurk workers might also know your procedure. For example, Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak, and Greene (2014) used a popular paradigm, the public goods game, to examine cooperation. They found that inexperienced mTurk workers tended to cooperate. Experienced mTurk workers, who presumably knew that they would get more points by being selfish based on past experience playing the game, cooperated less.

If you are worried that mTurk workers are already familiar with the study you want to run, you have options. The most drastic is to run it online with a subject pool. But there are also ways to screen and exclude participants based on their prior experience, as we discuss in the section on *Insuring that the same person does not do your study multiple times*.

Insuring the same person does not complete your study multiple times

In online data collection, there is an inherent risk that one person could complete a particular study multiple times. There are two types of participants to worry about. One is the participants we talked about in the previous section, those who are non-naive because they participated in a HIT posted by someone else. First, though, we will outline ways to avoid having the same participant do more than one of your HITs. One way to minimize the risk of having the same participant complete your study multiple times is to add assignments to the same HIT. Each unique mTurk ID is only able to complete any given HIT one time. (Note: Be careful about adding a new batch to a HIT, because if you do that, a participant from batch one can also participate in batch two. Adding assignments to a single batch is a better strategy.) There are drawbacks to this approach, however: We have found that participants are slower to sign up for an added assignment than for the initial assignment (probably because added assignments do not make it to the top of mTurk's HIT list). Moreover, you cannot change a HIT's parameters when adding assignments (i.e., how long the HIT lasts, the mTurk instructions, the link in mTurk, or how much participants earn by completing the HIT). Excluding mTurk workers via qualifications does not have these problems. A filter can be applied on Mechanical Turk that excludes workers from completing a HIT if they have completed a prior HIT that you specify.

Check your own database of previous users

Some researchers maintain a database of previous users to make sure a participant does not participate in two studies that are too similar to each other. This system works as follows: A participant clicks a link on mTurk, arrives at your data collection website and enters their mTurk ID. You then check that ID against a database. If the ID has been used in a study that is incompatible with the study you are doing, the participant can be told that he or she is ineligible to complete the current study prior to starting it.

Next we turn to a more difficult problem: How to identify mTurk workers who are not naive because they have participated in someone else's HIT. You will not find these workers' IDs in your data and you cannot use Mechanical Turk filters to avoid them. There are two things that can be done, although admittedly neither is perfect.

Create a shared database of previous users

If you know of other labs that run participants online and use the same paradigm as you do, you might be able to create a shared database of users. You would do the same thing as you do when you check a database of previous users (described above), but multiple labs would agree on, contribute to, and check a shared database.

Always ask about prior participation

At the end of each experiment, it is good practice to ask participants if they have ever completed a study using the same materials before. If they say yes, you will have a record of it and can exclude them from the analyses, if the analyses call for it. It is important to communicate that answering 'yes' will not affect their payment, and that it is important for them to be honest.

We have no way of knowing how honest people are when answering this question, but we do find that people are willing to answer in the affirmative, provided that they know they will still be compensated for their work.

Excluding data

We just described the importance of excluding participants who have done your study before. There are a variety of other reasons to exclude data from your analyses.

Most of the factors that will affect your decision to exclude are not unique to online studies. For example, if a participant's recorded reaction times show that he or she did not take the time to read your instructions (you should record reaction times for instruction pages) or to engage with your task, you might want to exclude this participant. To avoid misunderstandings that lead to exclusion, write your instructions as clearly as possible. There are some online-specific issues, which we turn to next after making a general recommendation.

A general recommendation

We recommend trying to exclude as little data as possible. Differential attrition can invalidate the conclusions of a study, and even if attrition rates are equal across conditions, attrition can lead to a biased sample (see Chandler et al., 2014; Zhou & Fishbach, 2016). We have four recommendations. If possible, plan your study in a way that will not require a lot of exclusion. Decide on a set of exclusion rules before analyzing your data. When in doubt, be inclusive. And be transparent when reporting attrition and exclusion when reporting the data.

Excluding participants who do your study twice in a row

Our participants sometimes complete our entire study, or sometimes a just a few trials, more than once. Figuring out what they did is relatively easy to do, given that saved data files, and often individual trials, are timestamped. Sometimes the data are salvageable, for example when the participant completed the entire study before starting over, or when they repeated a single item that can be excluded. The participant's data usually need to be fully excluded if they do part of the study and then start over.

Participants who decided not to finish

In studies run in a lab or online with a subject pool, participants usually finish the studies they start. This might be because they sign up ahead of time, use their real names, or have a limited number of studies they can complete to get credit. On mTurk, though, we find that participants are more likely to stop part way through the study, often near the beginning. This kind of attrition can lead to a biased sample (e.g., a sample that consists only of participants who do not give up on unpleasant tasks) and should be considered with care.

Common myths about mTurk data collection

Myth #1: mTurk workers do not take experiments seriously

There is a large amount of evidence to suggest that mTurk workers take experiments seriously. Hauser and Schwarz (2008) conducted an online study that contained an instructional manipulation check (IMC). During the study, participants read the instructions and then were asked the question: "Which of these activities do you engage in regularly?" This question was followed by sports response options; however, participants were informed in the initial instructions to ignore all the response options. Instead, participants were instructed to type "I read the instructions" in a box labeled "other." The mTurk workers passed the IMC at a remarkable 95% rate. In contrast, college students completing the study online in exchange for course credit at a large Midwestern university had a pass rate of just 39%. These mTurk workers were paid 30 US cents to complete this survey, ruling out a strong financial incentive to do well on the task. Similar results were obtained in Paolacci, Chandler, and Ipeirotis (2010). These researchers had participants complete a variety of tasks, one of which was to answer the question: “While watching the television, have you ever had a fatal heart attack?” Obviously, any response other than “Never” is incorrect. Participants answering this question on mTurk had a numerically lower (albeit nonsignificant) failing rate (4.17%) compared to participants from a Midwestern university (6.47%). Simons and Chabris (2012) had mTurk workers answer general questions about memory (e.g., “People suffering from amnesia typically cannot recall their own name or identity”). Most laypersons endorsed these false statements as true. These researchers also conducted a more labor-intensive and expensive phone survey. When they analyzed the data after controlling for demographic differences, the inexpensive mTurk sample’s results matched those of the expensive phone sample. These experiments suggest that mTurk workers are generally attentive and take online experiments seriously—perhaps more seriously than a subject pool sample.

Myth #2: Studies conducted on mTurk do not replicate lab-based findings

There is a variety of evidence to suggest that mTurk studies do replicate lab-based findings. Berinsky, Huber, and Lenz (2012) conducted the “Asian Disease Problem” (see Tversky and Kahneman, 1981) on mTurk. In the experiment, participants must choose between two treatment options. Even though both options are mathematically identical in terms of number of lives that could be lost, there are clear biases in which option participants select depending upon the framing of the question and the alternatives presented. The results from the mTurk sample were almost identical to those originally observed by Tversky and Kahneman (1981).

Verkoeijen and Bouwmeester (2014) conducted a replication of Kornell and Bjork (2008) with an mTurk sample. Kornell and Bjork (2008) showed participants 6 different paintings from 12 different artists (e.g., Georges Braque). The paintings from each artist were either presented via a massed practice schedule (e.g., all 6 paintings by Georges Braque were studied consecutively) or via a spaced practice schedule (e.g., a particular artist's paintings were spaced across numerous blocks in the experiment). In Experiment 2 a final recognition test was given during which participants were presented with novel paintings that had not been studied initially and had to indicate which artist painted that particular painting. Kornell and Bjork (2008) found that presenting the paintings via a spaced schedule resulted in the best performance in terms of categorizing the novel paintings on the final test. Furthermore, the majority of participants indicated that they had learned the most during the massed study schedule. Verkoeijen and Bouwmeester (2014) conducted a highly similar experiment on mTurk and achieved the same pattern with respect to final test performance and the participants' metacognitive judgments. Additionally, the effect sizes obtained by Verkoeijen and Bouwmeester (2014) were remarkably similar to those obtained by Kornell and Bjork (2008).

Other research supports these findings. Casler, Bickel, and Hackett (2013) showed participants two tools. One of these tools was an ordinary, familiar object performing its usual function (e.g., a paintbrush painting something). In contrast, the other tool was similar in shape and general appearance to the familiar tool but was novel and unknown. In the laboratory version, participants physically held and manipulated the novel tool. In the mTurk version, participants watched videos of the novel tool to get a sense of its function and usage. For both versions of the experiment, there were "teaching trials" and "non-teaching trials". On "teaching trials," the researchers demonstrated how the novel tool was used (i.e., it was shown performing an action). On "non-teaching trials," only physical descriptions were used when discussing the novel tool (e.g., color). Next, a task was presented to the participants, and participants had to select which tool they would use to complete the task. Although each tool could sufficiently complete the task, participants selected the novel tool more often during “non-teaching trials”. This pattern obtained not only in the laboratory, where participants could physically touch and handle the tools, but it also appeared in an mTurk sample. Overall, these findings suggest that online studies can produce results similar to lab-based studies.

Myth #3: Only simple experiments can be conducted online

Although surveys and simple tasks are easily implemented within mTurk, this and other crowdsourcing platforms are capable of much more than that. Staffelbach et al., (2014) had mTurk workers complete a complex citizen-engineering task. After reading a tutorial, mTurk workers were asked questions related to Virtual Wind Tunnel data analysis. mTurk workers had to interpret graphical simulations to determine if the simulation should be kept. The results suggested that mTurk workers can learn to complete tasks requiring basic engineering analysis and were slightly more accurate than trained graduate students.

Marge, Banerjee, and Rudnicky (2010) had mTurk workers listen to audio recordings and transcribe them. Workers were paid either $.005, $.01, $.03, or $.05 per transcription. The quality of the transcriptions were calculated using WER (word error rate). WER percentages with the mTurk sample were similar to audio transcriptions completed in-house. Furthermore, WER percentages were similar across payment amounts, providing further evidence that mTurk quality does not necessarily depend upon how much the workers are paid.

Hornsby and Love (2014) had mTurk workers view mammograms and attempt to classify them as either normal or tumorous. During the acquisition phase, mTurk workers were shown either exclusively easy mammograms (i.e., either very clearly normal or tumorous) or a range of mammograms that varied in terms of classification difficulty. The results indicated that mTurk workers trained on the easier mammograms were better at classifying novel mammograms later. This is an example of a memory study using complex stimuli.

In contrast to these studies, memory research often uses relatively simple stimuli. But there are other complications. For example, laboratory-based memory research is plagued by “problems of convenience”, one of which is using very short retention intervals (i.e., the delay between study and test). For instance, Cepeda, Pashler, Vul, Wixted, and Rohrer (2006) conducted a literature review of how *distributed practice* (i.e., spacing out practice, as opposed to *massed practice*) influences recall. Out of a possible 254 studies, only 1 study used a retention interval of 31 days or longer, and only 6 studies used a retention interval between 8-30 days (see Table 1 on p. 359). The majority of these studies used a retention interval between 1 second – 10 minutes (83.86%; see Table 1 on p. 359). Implementing longer delays is difficult to do in the laboratory, but easy to do online. Participants can be sent a reminder email at any time, and can complete the next phase of the experiment anywhere with an internet connection.

These findings suggest that mTurk workers are capable of learning the basics of civil engineering, can successfully transcribe audio recordings, and can even learn to classify mammograms as normal or tumorous. Clearly, mTurk is capable of conducting complex experiments, and is only limited by your creativity.

Case studies

As the studies above ought to suggest, online experiments are not limited to survey/questionnaire-based experimentation. In this section we describe innovative case studies that give a sense of the range of possibilities offered by online studies. In our search for innovative studies, we did not limit ourselves to research on memory, although that is the topic of this book. Instead, we selected studies with interesting methodologies that could easily be adapted for use by memory researchers.

Studying the effects of thirst on valuation of a novel drink. Haggag & Pope (in prep) were interested in whether a participant's state at the time of a novel consumption experience would influence their subsequent evaluation of it. They requested that participants retrieve a series of common ingredients from their kitchens--water, orange juice, milk, and sugar--and arrange the ingredients in a particular configuration alongside a handwritten card that included their worker ID. After arranging them, they were asked to take a picture with their camera and upload it (to ensure that participants complied with instructions). Next, they had participants make themselves a novel drink by mixing the orange juice, milk and sugar and then consuming it. To manipulate their current state at the time they experienced the novel drink, they had participants drink either ½ cup or 3 cups of water immediately before consuming their concoction, thus quenching the thirst of some participants but not others. Several days later, they contacted participants and asked them to complete a follow-up survey assessing their thoughts about the drink they consumed. Although this was not a memory study, it is easy to imagine a situation in which one would want participants to do some relatively complex task as part of an online memory study. For example, participants might be asked to look through an old photo album or yearbook, or they might be asked to move to a new room before studying each of a set of word lists. This case study suggests that doing experiments like this is possible online.

Increasing participant interaction. Another common strategy that can allow for more indirect interaction between participants involves having the responses from one group of participants serve as inputs to the experiences of a subsequent group of participants. A participant might make an offer to which a different participant subsequently responds, or the responses to a test completed by a set of test-takers can then be provided to a subsequent set of participants who can then grade them. In the domain of memory, these strategies would allow researchers to run yoked control conditions and make metacognitive judgments about others' memories, among other things.

Coding / norming data with mTurk. MTurk workers can even be tasked with the various roles of a research assistant. They can be asked to code the anonymous responses of other sets of MTurk workers, or they can pre-test materials on which they can be asked to provide feedback.

As all of these examples suggest, researchers are really only limited by their own creativity in devising clever ways of eliciting ecologically valid behaviors from online participants (within the bounds of what technology can do and their own technical abilities). In some ways, as we outline in more below, having access to participants completing tasks from the comfort of their own homes rather than in the sterile confines of a laboratory can help to create more ecologically valid behaviors.

Benefits for psychology

Running well-powered studies. Previous research has suggested that a large amount of psychological research is statistically underpowered. Simmons, Nelson and Simonsohn (2011) have demonstrated that small samples combined with flexibility in research design and analysis can dramatically increase the likelihood of false positives and unreliable findings. Mechanical Turk offers an opportunity to improve the reliability of our science by allowing researchers to gain access to a much larger pool of potential subjects than is typically available to the average researcher (though it is not limitless; see Stewart et al., 2015). This easily allows for samples of fifty or hundred or more per cell, which can increase the power to detect modest effect sizes of the sort that typically occur in psychological research, including memory research.

Running more replication studies. Whereas a direct replication might be exceedingly timely and costly to run when using an undergraduate subject pool (especially if attempting to achieve n=100 per cell), researchers can much more quickly and easily establish the robustness of their findings using online samples where participants are much more plentiful. Replicating one's own study, at least the data collection aspect of it, can require literally a few minutes' work. This can allow researchers to ensure that their findings are not spurious before they publish them.

Reaching a more diverse subject pool. Researchers have often relied too heavily on undergraduate college samples in their research, and have often implicitly assumed that any findings discovered in these samples will successfully generalize to other populations. However, a number of researchers have pointed to the limitations and weaknesses of exclusively studying Western, Industrialized Educated, Rich, Democratic (WEIRD) samples (Henrich, Heine & Norenzayan, 2010). Though Mechanical Turk does not fully realize the goal of achieving a completely nationally representative sample, it is nonetheless much more diverse than a typical undergraduate sample, allowing researchers to study a much broader cross-section of ages, ethnicities, political leanings, and participants from different socioeconomic backgrounds (Berinksi et al., 2012).

Running more multi-session studies. Mechanical Turk is a relatively easy way to conduct longitudinal studies in which participants complete tasks over multiple sessions separate by potentially large amounts of time. Access to undergraduate samples often waxes and wanes with the ebb and flow of the semester, and, moreover, these participants stop being interested in participating in studies when the semester is over. By contrast, mTurk workers might use the service for many years. This allows researchers to establish longer-term relationships with requesters and provide data to follow-up experiments over much longer periods of time. Additionally, because multi-session studies are much less labor-intensive than multi-session lab studies, they afford greater opportunities to establish the durability of effects or explore the longer-term trajectories of them beyond a single session.

Increased robustness and external validity. Sitting in an unfamiliar laboratory can make people behave abnormally. A lab can be a strange environment and might make some participants uncomfortable, as might being in close proximity to authoritative researchers. Participants might become uptight and nervous. They might also lie, or at least shade the truth, to present themselves in a way that makes them look good. These concerns are all lessened in online studies in which participants are at home and anonymous.

Participants might be more on task and responsible in a lab than they would be in real life because they are not anonymous and they know they are being monitored. This has obvious advantages, but it has disadvantages as well. The most robust psychological effects, the ones that actually affect people in real life, should be detectable under normal life circumstances. That might mean that a participant does your study with a baby on one knee, has music playing in the background, and answers a few text messages during your study. If an effect cannot be detected under this kind of treatment, we argue that it might not actually affect people often in real life. In short, conducting studies online might be a way to insure that whatever effects to do obtain have more external validity and robustness than a study run in a lab.

Finally, running a study in the most realistic possible way means meeting participants where they are; to study schooling means conducting research in a school. Increasingly, “where they are” is online. Impression formation and personal interaction happen on social media websites and apps; multitasking happens when participants open another browser tab planning to come back to the first; students study using flashcard-like apps; games that require players to come back and do small tasks to earn rewards are all about operant conditioning. An online study that can examine behavior in these situations may have high external validity.

Closing remarks

With the arrival of Mechanical Turk, Qualtrics, Survey Monkey, Google Forms, and a host of other services and technologies, it has never been easier to reach online participants and recruit them for your studies. This allows for opportunities not only to efficiently collect large numbers of subjects, but also to study a much more diverse population than the average undergraduate subject pool and to conduct much higher-powered studies that can prevent faulty inferences drawn from your data. In the words of one of our colleagues, these advantages make online data-collection a “game changer” for psychology.

The explosion of interest in online data collection has happened concurrently with another evolution in Psychology: the recognition that many previous data collection and analysis strategies suffered from a variety of problems that led to false positives, weakened the accuracy of inferences drawn from data, and weakened the robustness of psychological phenomena. Given that both of these trends are unlikely to halt or reverse anytime soon, now is exactly the right time for you to conduct your first online study.

References

Bartels, L. M. (1999). Panel effects in the American national election studies. *Political Analysis*, *8*(1), 1-20.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3), 351-368.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156-2160.

Chabris, C. F., & Simons, D. J. (2010). *The Invisible Gorilla: And Other Ways Our Intuitions Deceive Us*. Broadway Paperbacks.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. http://doi.org/10.3758/s13428-013-0365-7

Christenson, D. P., & Glick, D. M. (2013). Crowdsourcing panel studies and real-time experiments in MTurk. *The Political Methodologist*, *20*(2), 27-32.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system?: screening mechanical turk workers. *Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Gustavson, K., von Soest, T., Karevold, E., & Røysamb, E. (2012). Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a Monte Carlo simulation study. *BMC Public Health*, *12*(1), 1-11.

Hauser, D. J., & Schwarz, N. (2016). Attentive MTurk workers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1), 400-407.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*(2), 72-76.

Kittur, A., Chi, E. H., & Suh, B. (2008, April). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453-456). ACM.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585–589.

Marder, J., & Fritz, M. (2015). The Internet's hidden science factory. *PBS Newshour*. Retrieved from http://www.pbs.org/newshour/updates/inside-amazons-hidden-science-factory/.

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). Using the amazon mechanical turk for transcription of spoken language. *Paper presented at the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.

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

Rand, D. G., Greene, J. D., & Nowak, M. A. (2013). Rand et al. reply. *Nature, 498*(7452), E2-E3. doi: 10.1038/nature12195

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, http://doi.org/10.1038/ncomms4677

Rand, D. G., Greene, J. D. & Nowak, M. A. (2012) Spontaneous giving and calculated greed. *Nature, 489*, 427–430.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Simons, D. J., & Chabris, C. F. (2012). Common (mis) beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PloS one*, *7*(12), e51876.

Staffelbach, M., Sempolinski, P., Hachen, D., Kareem, A., Kijewski-Correa, T., Thain, D., Wei, D. and Madey, G. (2014). Lessons learned from an experiment in crowdsourcing complex citizen engineering tasks with Amazon Mechanical Turk. *arXiv preprint arXiv:1406.7588*.

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, *10*(5), 479-491.

Stoycheff, E. (2016). Please participate in Part 2: Maximizing response rates in longitudinal MTurk designs. *Methodological Innovations*, *9*, 2059799116672879.

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., Västfjäll, D., Kirchler, M., & Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature*, *498*(7452), E1-E2.

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*, 453-458.

Zhou, H., & Fishbach, A. (2016). The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/ pspa0000056.

Appendix A

Links to websites involved in online data collection

Service	Link
Amazon’s Mechanical Turk	www.mturk.com
Collector	https://github.com/gikeymarcia/Collector
Google Forms	https://www.google.com/forms/
Inquisit by Millisecond	www.millisecond.com
psiTurk	https://psiturk.org/
Qualtrics	https://www.qualtrics.com/
Survey Monkey	https://www.surveymonkey.com/
Turkopticon	https://turkopticon.ucsd.edu/