The (Statistical) Power of Mechanical Turk

Amelia Kimball*
University of Illinois at Urbana-Champaign

In this paper, I argue for the use of Amazon Mechanical Turk (AMT) in language research. AMT is an online marketplace of paid workers who may be used as subjects, which can greatly increase the statistical power of studies quickly and with minimal funding. I will show that—despite some obvious limitations of using distant subjects—properly designed experiments completed on AMT are trustworthy, cheap, and much faster than traditional face-to-face data collection. Not only this, but AMT workers may help with data analysis, which can greatly increase the scope of research that one researcher may carry out. This paper will first argue several reasons for using online subjects, then quickly outline how to build a survey-type experiment using AMT, and finally review several best practices for ensuring reliable data.

Keywords: Mechanical Turk, online experiments, big data, statistical power.

1 Introduction

This paper argues for the widespread use of Amazon Mechanical Turk (AMT) in language research. AMT is an online marketplace run by Amazon.com. On the marketplace, requesters can post Human Intelligence Tasks (HITs) for online workers to do. Requester posts include a description of the task and the amount of money that the worker will be paid. Requesters pre-pay Amazon for the HITs, workers do the HITs and submit them for approval, and then requesters either approve the work or reject it. Workers receive money for completed and approved HITs, and Amazon charges requesters a 10% fee. This marketplace is widely used by businesses to find participants to do short, simple, tasks that are better accomplished by humans than by computers, such as picture tagging or reading scanned documents. Recently, AMT has become popular as a means of recruiting subjects and collecting data in the social sciences, though it has yet to establish itself as the default form of data collection.

Many studies have investigated the trustworthiness of data collected in online social science experiments. As a group, these studies have found data to be trustworthy. Specifically, they found (1) that agreement in text annotations between groups of inexperienced workers and individual expert annotators (Snow et al. 2008); (2) that workers are generally truthful about
information such as their location (Ipeirotis 2011); and (3) that workers’ responses to psychometric, judgment and decision-making tests are as reliable as those collected with traditional methods (Buhrmester, Kwang & Gosling 2011; Paolacci, Chandler & Ipeirotis 2010). Despite this evidence, some researchers remain unconvinced that online methods are useful or even valid. In this paper, I will argue that online data collection is not just as good as traditional methods, but in some cases even better. I will first outline a four-point argument as to why using AMT for data collection is advantageous for language researchers, with some limitations (Section 2). Then, I will present a brief overview of how AMT experiments may be built and few best practices for doing online studies (Section 3).

2 Why Use Data from Online Subjects?

It may be no surprise to the reader that online experiments are in many senses more convenient than traditional laboratory studies, and I will outline the reasons why this is the case in sections 2.1-2.3. However, the point that I wish to underline in this paper is that this convenience is not just a boon to busy researchers, it also allows linguists to build better studies that are accessible to bigger populations of researchers and subjects (section 2.4). In brief, data from AMT is useful because compared to traditional face-to-face data collection it’s faster, it’s cheaper, it’s more varied, and there’s more of it. In many cases this has the potential to create better, more statistically valid studies.

2.1 Faster

The first undeniable advantage to AMT studies is the speed of recruiting participants and collecting data. Researchers will find that studies posted to AMT are often completed in hours, rather than days or weeks of lab time. Researchers can also use the pressures of the marketplace to their advantage—the higher paying the study, the faster it will be completed, and vice versa (Buhrmester, Kwang & Gosling 2011). In many cases, it is as easy to build an experiment in an online survey platform as it would using dedicated software for lab use. The great advantage to the online versions is that once built, the experiment can be run by many subjects simultaneously. What’s more, due to time zone differences, the window where target subjects are awake and working is much larger than a 9-5 workday. Additionally, because many thousands of workers are online at any moment, finding many subjects—even many members of a special population—takes very little time and requires no time spent scheduling subjects. Many experiments will need little supervision from the experimenter, freeing up hours of experimenter time for other parts of the research cycle.

2.2 Cheaper

Typical studies on the University of Illinois Urbana Champaign campus in 2014 paid approximately $8.00 per hour. Because paperwork and equipment set up may take time, and because subjects are harder to recruit for shorter lengths of time and lower pay rates, many language researchers at UIUC end up paying $8 per subject or more. In contrast, typical HITs
on AMT pay very little—in a survey from 2010, Ipeirotis reported that 90% paid less than 10 cents per subject. This is partially due to the fact that many HITs are short—only seconds long in many cases. Because workers are already online and are used to doing many low-paying HITs in a work session, they are very willing to perform longer HITs (i.e. > 2 minutes) that are relatively high-paying (>$1/ HIT). This means that researchers can choose to lower the hourly wage that they pay. Or, they may choose to maintain a high hourly wage while significantly dropping their per subject cost. This can be achieved by designing studies that take less than an hour and eliminate setup and paperwork time due to their online nature. Of course, researchers should act ethically and within the bounds of approval from their institutional research board and state and federal wage limits, but many will find that there is more flexibility within these limits in online studies than in face-to-face research.

2.3 More Varied

Because the population of AMT workers is so much larger and more widely distributed geographically than the population of any university, it is more varied than the sample available to most researchers. (See the survey reported in Ipeirotis (2010) for an overview of demographic information). As of this writing, Amazon pays in only US Dollars and Indian Rupees, so populations are strongly skewed toward American and Indian workers. However, this being said, there are workers from many countries and many different language backgrounds, not to mention workers of many different ages (for an overview of language backgrounds see Pavlick et al. 2014). Any researchers whose local sample of convenience is restricted in age or language background will benefit from access to this worldwide pool of participants (for more on targeting special populations see section 3.3 below). This is important not just because researchers may look for special populations, but also because a sample of AMT workers will be more varied and in that sense more representative of the underlying population. In other words, the students of a university are already themselves a “special population”—often with a narrow age range and relatively narrow socio-economic status. If the goal of an experiment is to test a hypothesis in, say, the entire population of American English speakers across ages, genders, and locations, a sample of AMT workers will be closer to a representative sample than an equivalently large group recruited in the hall of the foreign languages building at a university.

2.4 There’s More of It

The fact that many AMT studies are faster and cheaper has instant appeal for the busy graduate student or the underfunded program. However, the point that I wish to underline in this section is that the sheer volume of data made possible by this convenience also makes for more statistically valid studies.

At the outset of any experiment using subjects, researchers must determine the number of subjects that will let them test their hypothesis in a statistically valid way. Ideally, when determining the number of subjects to use, researchers would avoid underpowered or overpowered studies by estimating effect size and conducting a statistical power analysis to determine the proper number of subjects. In practice, this may be difficult because the number
of participants may be restricted by logistical constraints, and effect size may be hard to estimate. Additionally, I acknowledge that very few linguistic studies run enough participants to create a representative sample of the underlying population they are studying, since the underlying population may be massive (e.g. “speakers of French”). However, when effect size is hard to estimate and the experiment has no risk for participants, the general rule is that more data is better, since it will increase the statistical power of the study no matter what the effect size. With AMT, researchers can get far more participants faster and cheaper and increase the statistical power of their studies—and therefore the reliability of their results.

2.5 Caveats and Downsides

Of course, AMT is not perfect for every experiment. It is obvious that some techniques such as eyetracking or fMRI can never be performed with distant participants. Additionally, perhaps the biggest drawback of AMT is that the experimenter has no control over the participants’ environment—there is nothing stopping participants from performing the experiment in noisy or distracting conditions. Experimenters that are used to presenting stimuli in controlled environments such as sound attenuated booths or purposefully bare experiment rooms will have to prepare experiments that are robust to unknown amounts of real world distraction. Experiments that target very small effect sizes and require intense attention from participants are not well suited to AMT. That said, larger numbers of participants can often mitigate this downside, and at the end of the day many researchers are interested in language in distraction-heavy real world environments, rather than in highly controlled situations.

Another drawback of AMT is the potential for a higher attrition rate—more subjects may choose to not finish the experiment. Logically, it is easier for a participant close a browser window than to quit and leave a face-to-face session. This too is typically mitigated by larger numbers of participants and the fact that only participants who finish the task are paid. Lastly, there is always the chance that participants can lie about their demographic information, in ways that might not be feasible in a face-to-face interaction. With less interaction comes less pressure on subjects to be truthful. Overall, each experimenter must choose what technique is best for them. An overview of advantages and disadvantages is presented below in figure 1.
3 Using AMT

This section outlines what types of experiments are feasible on AMT and how to build a survey type HIT.

3.1 What Can You Use AMT For?

AMT can be used for any experiment that can be built as a website. The most obvious and easy of these is to simply design a survey using some of the many survey-building tools available on the Internet (e.g. SurveyGizmo, SurveyMonkey, Qualtrics; see references for URLs). Though giving surveys may bring to mind the realm of social psychology, many language experiments at their heart involve simply asking participants a series of questions:

- Are these sound files the same or different? (AX, ABX etc. tasks)
- Do these sentences sound natural to you? (grammaticality judgments)
- How would you finish this sentence? (sentence completion tasks)
- Do these words mean the same thing? (semantic judgments)
- Would you say this to your boss? Your friend? (pragmatic judgments)
- Who do you think said this? (sociolinguistic matched guise tasks)

These types of surveys are useful in all subfields of linguistics. Even the most theoretical parts of linguistic inquiry must be grounded in speaker intuitions, and therefore even the most theoretical of linguists can benefit from online experiments.

Surveys are the most user-friendly to build because Many free, user-friendly tools exist for building them. However, many other options exist to those with programming experience. Most

---

Figure 1: Advantages and disadvantages of recruitment techniques.
experiments that involve a subject sitting at a computer can be implemented online using open source software such as Ibex and PsychoPy (see references). For example, self-paced reading tasks, behavioral experiments involving collection of reaction times, and mouse tracking experiments can all be implemented online. Subjects can interact through purpose-built chat rooms or computer game interfaces. Subjects can also make recordings using web interfaces, though the recordings may be noisy due to poor recording equipment on subjects’ computers. In essence, almost any experiment that does not require a subject to be physically present in the lab is possible to implement on AMT.

Researchers should also keep in mind that AMT can be of use in all parts of the research cycle, not just for data collection itself. Workers may help to norm stimuli, to pilot experiments before wider release, or to annotate or judge responses in data analysis.

3.2 How to Build a Survey Experiment in AMT

Within the AMT interface researchers can create and publish HITs, including asking survey-type questions. However, for surveys that are longer than a few questions, many researchers choose to use an outside survey program. Typically, survey programs have the advantages of being easy to use and of providing reports summarizing trends in participant responses. To build a HIT on AMT that uses a survey built with an outside survey tool, simply create a “Survey link” HIT from the survey link HIT template, and then provide the URL of your survey to participants on AMT. You may choose to give the workers a code at the end of the outside survey. This code is then entered on AMT, so that you can quickly verify from the AMT interface that the worker has finished the survey before paying her or him. Be sure to test your survey before you make it live. Test it in all popular web browsers, particularly to make sure that media such as sound files and pictures appear correctly. For testing within the AMT interface itself, especially for more complicated HITs, you may choose to use the “developers sandbox.” This is an identical version of the AMT site that is not live to workers, so that you can test functionality before publishing your HITs. An overview of the process of building a survey HIT is in figure 2 below.
3.3 Targeting Special Populations

If your aim is to target a specific population, there are several ways to identify certain demographics. The simplest, but most vulnerable to worker deception, is to ask in your advertisement for specific workers and hope that they are the only ones that respond. Alternatively, you can post a short demographic survey as a HIT, and then contact only those workers who are part of the demographic you are interested in. Additionally, AMT allows you to restrict your HIT to only workers in a certain country, which can be helpful when targeting a specific language (variety). If the population you are interested in is widespread, you can simply advertise and pay all participants, and only use data from the population you are interested in. Some are averse to “wasting” data this way, but this ensures worker honesty because workers have no incentive to lie, and HITs are typically cheap enough that generating unusable data will be much more bearable than in face-to-face collection.

3.4 Optimizing Data Reliability

When building an experiment on AMT, there are several ways to ensure that data is maximally reliable. Here I outline four suggestions for optimizing reliability.

Firstly, as hinted at above, create no incentive to lie. Avoid telling subjects that your study is limited to certain demographics, which may motivate them to stretch the truth in order to be eligible to participate. Instead, use common sense ways of filtering subjects with pre-screen
surveys, built-in AMT tools, and smart tricks. If you need fluent Russian speakers, try writing a grammatically complex task in Russian.

Second, maximize participant attention by building attention checks into your design: timestamp all parts of your experiment, collect worker IDs, and check those IDs against the results of your experiment. You may also ask workers to identify if a picture or audio does not load, and then purposefully disable some of the pictures/audio to check that the participant is paying attention.

Third, set a priori bounds for acceptable data (in terms of participant background, amount of time spent on a trial, accuracy, etc.). Budget for data loss and be ruthlessly conservative when it comes to throwing away data that does not meet those standards.

Lastly, create a reputation as a good requester. Pay workers quickly and generously. This will aid your reputation on AMT. Workers share reviews of requesters, and those who are reported to be fair and generous will get their studies done the quickest and by the most experienced workers. If a participant has obviously not followed directions, reject their work but explain clearly your reasons for rejecting it. However, err on the side of paying all participants generously, even ones who do not follow directions—chances are your reputation as a generous requester is worth more than accidentally overpaying a subject that did not correctly complete the task.

4 Conclusion

This paper has argued that data collected on AMT is cheaper, faster, and more varied than that collected by traditional methods. The convenience of collecting large amounts of data on AMT makes for better studies with higher statistical power. Though AMT is not appropriate for every language study, researchers who are under time or funding pressure should make efforts to see if their studies can be implemented using AMT, in order obtain faster, cheaper, more reliable results.

References


