Crowdsourcing Samples in Cognitive Science

Neil Stewart, Jesse Chandler, Gabriele Paolacci

June 8, 2017

Abstract

Crowdsourcing data collection with research participants from online labor markets is now well established in cognitive science. We review who is in the crowd and who can be reached by the average laboratory. We discuss reproducibility and review some recent methodological innovations for online experiments. We consider the design of research studies and arising ethical issues. We review how to code experiments for the Web, and what is known about video and audio presentation and the measurement of reaction times. We close with comments about the high levels of experience of many participants and an arising tragedy of the commons.

1 Data Collection Marketplaces

Online labor markets have become enormously popular as a source of research participants in the cognitive sciences. These marketplaces match researchers with participants who complete research studies in exchange for money. Originally, these marketplaces fulfilled the demand for large-scale data generation, cleaning, and transformation jobs that require human intelligence. However, one such marketplace, Amazon Mechanical Turk (MTurk), became a popular source of convenience samples for cognitive science research. Based upon searches of the full article text, we find that in 2017, 24% of articles in Cognition, 29% of articles in Cognitive Psychology, 31% of articles in Cognitive Science, and 11% of articles in Journal of Experimental Psychology: Learning, Memory, and Cognition mention "MTurk" or another marketplace, all up from 5% or fewer five years ago.

Box: MTurk Terms

MTurk. The Amazon Mechanical Turk labor marketplace
HIT (Human Intelligence Task). A task that is posted on MTurk by a requester for completion by a worker
Workers. People who subscribe to MTurk to complete HITs in exchange for compensation
Requesters. People or companies that post HITs on MTurk for workers to complete

Reward. The compensation that is promised to Workers who successfully complete a HIT
Qualifications. Requirements that a Requester sets for Workers to be eligible to complete a given HIT. Some qualifications are assigned by Amazon and available to all requesters. Requesters can also create their own qualifications
Approval/Rejection. Once a Worker completes a HIT, a Requester can choose whether to Approve the HIT (and compensate the worker with the reward) or Reject the HIT (and not compensate the worker)
HIT Approval Ratio. The ratio between the number of approved tasks and the number of total tasks completed by a worker in her history. Until a worker completes 100 HITS, this is set at a 100% approval ratio on MTurk
Block. A requester can "block" workers and disqualify them from any future task they post. Workers are banned from MTurk after an unspecified number of blocks
Command Line Tools. A set of command line interfaces allowing the Requester to access more features of MTurk than via the user interface
Worker file. A comma separated values (CSV) file downloadable by the Requester with the list of all the workers who completed at least one task for the Requester
Turkopticon. A website where workers rate Requesters based on several criteria
TurkPrime. One among the websites that augments or automates the use of several MTurk features [1]

Although researchers have used the Internet to collect convenience samples for decades [2], MTurk allowed an exponential-like growth of online experimentation by solving several logistical problems inherent to recruiting research participants online. Marketplaces aggregate working opportunities and workers, ensuring that there is a critical mass of available participants, which increases the speed at which data can be collected. Moreover, these marketplaces offer rudimentary reputation systems that incentivize conscientious participation, and a secure payment infrastructure that simplifies participant compensation. Though marketing research companies provide similar services, they are often prohibitively expensive [3, 4]. Online labor markets allow participants to be recruited directly for a fraction of the cost.

Academic interest in MTurk began in computer science, which used crowdsourcing to perform human intelligence tasks such as transcription, developing training data sets, validating machine learning algorithms, and engage in human factors research [5]. From there, MTurk diffused to subdisciplines of psychology such as decision making [6], and personality and social psychology [7]. Use of Mechanical Turk then radiated out to other social science disciplines such as economics [8], political science [3] and sociology [9] and applied fields such as clinical science [10], marketing [11], accounting [12] and management [13].

Across fields, research using online labor markets has typically taken the form of short, one-shot surveys conducted across the entire crowd population or on a specific geographic subpopulation (e.g., US residents). However, online labor markets are quite flexible, and allow creative and sophisticated research methods (see Box "Innovative MTurk Research") limited only by researchers' imaginations, and their willingness to learn basic computer programming (see Box "Coding for Crowdsourcing"). Recently, a number of new tools (e.g., TurkPrime, which facilitates using more advanced MTurk features) and competitor

marketplaces have lowered technical barriers substantially.

Box: Innovative MTurk Research

Infant attention. Parents have been recruited on MTurk and their infants' gaze recorded using a webcam whilst the infants viewed video clips [14, 15]. Some participants were excluded because of technical issues (e.g., the webcam recording becoming desynchronized from the video playing) or because the webcam recording was too poor quality to view the infant's eyes. But it was possible to identify the features of videos that attracted infant attention, including singing, camera zooms, and faces.

Economic games. Many economic games have been run on MTurk, including social dilemma games and the prisoner's dilemma [16]. The innovation here was to have remotely located participants playing against one another live, where one person's decision affects the outcome another receives, and vice versa. Software platforms to implement economic games include Lioness [17] and NodeGame [18].

Crowd creativity. Several researchers have assembled groups of workers to engage in iterative creative tasks [19]. More recently, workers have collaboratively written short stories, and with changes in the task structure influencing the end results [20]. The innovation here was using microtasks to decompose tasks into smaller subtasks and explore how changes to the structure of these tasks changes group output.

Transactive crowds. A number of researchers have looked at how crowds can be used to supplement or replace individual cognitive abilities. For example, MTurk workers have provided cognitive reappraisals in response to the negative thoughts of other workers [21], and an app has been developed to allow people with visual impairments to upload images and receive near real-time descriptions of their contents [22]. The innovation here was allowing workers to engage in tasks in contexts that are unsupervised, have little structure and occur in near real-time.

Workers as Field Researchers. Meier et al. [23] had participants take pictures of their thermostats and upload them to determine whether they were set correctly. The innovation here was using workers to collect field data about their local environment and relay it back to the researchers.

Mechanical Diary. Mechanical Turk allows researchers to recontact workers multiple times, allowing for longitudinal research. Boynton and Richman [24] used this functionality to conduct a two week daily diary study of alcohol consumption.

Time of Day. It is also easier to test workers at different times of day, without the disruption of a visit to the laboratory. Dorrian et al. [25] found rightward shifts in spatial attention from peak to off-peak times of day for morning but not evening types.

Crowds as Research Assistants. Online labor markets were originally created to facilitate human computation tasks like content coding for which academics have traditionally relied upon research assistants. Crowds often produce responses that are equivalent or superior to the judgments of experts. For example, Benoit et al. [26] used MTurk to rate the ideology of statements from political texts: they found that 15 workers produce

ratings of equivalent quality to 5 political science PhD students and faculty. Importantly, crowds can return data exceptionally quickly: workers were able to content code 22,000 sentences in under five hours for $360.

Box: Coding for Crowdsourcing

MTurk is accessible through a graphical user interface (GUI) that can perform most of the platform functions. MTurk also provides an API that can be accessed via Python [27] and can simplify tasks such as contacting workers in bulk or assigning variable bonuses. More recently, TurkPrime has developed an alternative GUI that offers an extended set of features.

It is possible to code and field simple survey experiments entirely within the MTurk platform using HTML5, but functionality is quite limited. Most researchers create a HIT with a link to an externally hosted survey and a text box in which the worker can paste a confirmation code upon completing the study. Simple surveys might be conducted using a variety of online platforms such as Google forms or www.surveymonkey.co.uk, which require no programing skills. Surveys with more complex designs (e.g., complex randomization of blocks and items) may benefit from using Qualtrics, which also requires minimal programming skill. Researchers can also direct users to specialized platforms designed to program reaction time experiments or group interactions.

Researchers with complex designs or who wish to include a high degree of customization can program their own surveys. Early Web experiments were often coded in Java or Flash, but these languages are now largely obsolete, being unavailable on some platforms and switched off by default and having warning messages on others. Most web experiments are now developed using HTML5, JavaScript (which is not Java), and CSS—the three core technologies behind most Web pages. Broadly the HTML5 provides the content, the CSS provides the styling, and the JavaScript provides the interaction. Directly coding using these technologies provides considerable flexibility, allowing presentation of animations, video, and audio content [28, 29]. There are also advanced libraries and platforms to assist, including www.jsPsych.org which requires minimal programing [30], the online platform www.psiturk.org [31], www.psytoolkit.org which allows online or offline development [32, 33], and Flash-based scriptingRT [34].

Web pages are not displayed in the same way across all common browsers, and as yet not all browsers support all of the features of HTML5, JavaScript, and CSS. Further, the variety of browsers, browser versions, operating systems, hardware, and platforms (e.g., PC, tablet, phone) is considerable. It is certainly important to test your Web experiment across many platforms, especially if the exact details of presentation and timing are important. Libraries like jQuery can help overcome some of the cross-platform difficulties.

Box: Online Reaction Times

Many cognitive science experiments require accurate measurement of reaction times. Originally these were recorded using specialist hardware, but the advent of the PC allowed recording of millisecond-accurate reaction times. It is now possible to measure reaction times sufficiently accurately in web experiments using HTML5 and Javascript. Alternatively, MTurk now permits the downloading of software, which allows products like Inquisit Web to be used to record reaction times outside of the browser.

Reimers and Stewart [28] tested 20 different PCs with quite different processors and graphics cards, and a variety of MS Windows operating systems and browsers using the Black Box Toolkit. They compared tests of display duration and response timing using web experiments coded in Flash and HTML5. The variability in display and response times was mainly due to hardware and operating system differences, not Flash / HTML5 differences. All systems presented a stimulus intended to be 150 ms for too long, typically by 5–25 ms, but sometimes by 100 ms. And all systems overestimated response times by between 30–100 ms and had trial-to-trial variability with a standard deviation of 6–17 ms [see also 35]. If video and audio must be synchronised, this might be a problem. There are quite large stimulus onset asynchronies of about 40 ms across different hardware and browsers, with audio lagging video [29]. Results for older Macs are similar [36].

The measurement error added by running cognitive science experiments online is, perhaps surprisingly, not that important [37]. Reimers and Stewart simulated a between-participants experiment comparing two conditions with a known 50-ms effect size. Despite the considerable variability introduced by the (simulated) hardware differences across (simulated) participants, only 10% more participants are required to maintain the same power as in the counterfactual experiment with zero hardware bias and variability [28]. In the more usual within-participants experiment where the constant biasing of response times cancels in the difference between conditions, there is no effective loss from hardware differences (see also [38]). Accuracy is higher using the Web Audio API [29]. Reaction time data have been collected and compared in the lab and online, with similar results for lexical decision and word identification times, and Stroop, flanker, Simon, Posner cuing, visual search, and attentional blink experiments [39–42].

MTurk is by far the dominant player in academic research and has been extensively validated. However, early evaluations of competitor platforms such as Clickworker [43], Crowdworks [44], Crowdflower [45], Microworkers [43, 46, 47] and Prolific (which focuses on academic research specifically)[47] have been encouraging, and are plausible alternatives for researchers.

2 Who is in the Crowd?

Who is in the crowd we are sampling from? Crowds are defined by their openness to virtually anyone who wishes to participate [48]. Since crowds are not selected purposefully,

they do not represent any particular population and tend to vary in terms of national representation: MTurk consists mostly of Americans and Indians; Prolific and Clickworker have larger European populations; Microworkers claims a large Southeast Asian population [49] and CrowdWorks has a primarily Japanese population [44].

The most popular (and thus well-understood) crowdsourced population, US MTurk workers, is more diverse than college student samples or other online convenience samples [3, 50], although not representative of the US population as a whole. It is biased in ways that might be expected of Internet users. Direct comparisons of MTurk samples to representative samples suggest that workers tend to be younger, more educated, but report lower incomes and are more likely to be unemployed or underemployed. European- and Asian-Americans are overrepresented and Hispanics of all races and African Americans are underrepresented. Workers are also less religious and more liberal than the population as a whole. (For large scale comparisons to the US adult population see [51, 52]).

Though representativeness is typically discussed in terms of demographic differences, MTurk workers may also have a different psychological profile to that of respondents from other commonly used samples. Workers tend to score higher on learning goal orientation [53] and need for cognition [3]. They also display a cluster of interrelated personality and social-cognitive differences: workers report more social anxiety [10, 54] and are more introverted than both college students [53, 55, 56] and the population as a whole [51]. They are also less tolerant of physical and psychological discomfort than college students [54, 57, 58], and more neurotic than other samples [51, 53, 55, 56]. Workers also tend to score higher on traits associated with autism spectrum disorders, which tend to be correlated with the other dispositions mentioned here [59].

3 Sampling from the Crowd

Who is in your sample? Just as workers who opt into a crowd are not representative of any particular population, workers who complete any given study may not be representative of the worker population as a whole. Instead, they represent some subset of the worker population that decides to complete a given task. Sometimes large differences in sample composition are observed, even between studies with several thousand respondents [59]. Because of the differences, we make reporting recommendations (see Box "Crowdsourcing Checklist for Authors, Reviewers, and Editors"). A number of design features contribute to these differences. For example, researchers set (or forget to set) qualification requirements on MTurk such as worker nationality (which correlates with English language comprehension [55, 59]), or level of experience and worker reputation (which correlates with attentiveness [45]).

Other decisions that are not explicitly part of a HIT's design can inadvertently influence sample composition. Research studies are usually available on a first come first serve basis. Consequently, survey respondents will vary in characteristics such as personality, ethnicity, mobile phone use, and level of worker experience as a function of the time of day that a task is posted [51, 60]. Sample composition will also change as sample size increases, as HITs that request fewer workers will fill up more quickly. In particular, small samples tend to overrepresent the savviest workers, who take advantage of software and online communities to quickly find available work. Workers have self-organised into communities, and those

workers tend to earn more [61]. Small studies that do not restrict workers from completing more than one study contain a large proportion of non-unique workers (that is, will tend to attract the same group of workers [62]). Sometimes events beyond researcher control will impact the sample. Studies that are publicized on online forums (e.g., Reddit) may also return disproportionately young and male samples [63]. The workers who complete a survey early are also different from workers who complete later, leading samples to change across data collection [51, 60].

Box: Crowdsourcing Checklist for Authors, Reviewers, and Editors

Research studies need to adequately describe their methods, including the sample they use, in order to maximize the replicability of their results. We highlight features below that are unique or have increased relevance for MTurk. The list is intended to be suggestive not mandatory, but it is probably not sufficient simply to report that a study was "run on MTurk".

Collected automatically by MTurk or set by the experimenter:

- ☐ Qualifications required to take the HIT: country of residence, minimum HIT approval ratio or number of previously completed HITs, and any other custom qualifications. The sample that completes a HIT can differ dramatically depending on the qualifications used. For example, country of residence and HIT approval ratio have large effects on resulting data quality [47, 59]
- ☐ Properties of the HIT preview which affect which workers take the HIT: reward offered, any bonus pay and stated duration. Participants experience in the study begins with the decision to accept a HIT. If study materials are archived in a repository, a copy of the HIT title and text description should be included
- ☐ Actual duration of the HIT (perhaps median or range) and actual bonus pay
- ☐ Time of day and date for batches of HITs posted. Worker characteristics vary dynamically across time [51, 60]

Consider recording:

- ☐ Whether workers were prevented from repeating different studies in the same package of studies. Most traditional subject pools prevent repeated participation by default. Mechanical Turk does not, and repeated participation affects participant responses
- ☐ Attrition rates (because many people can preview the study before accepting the HIT). Attrition rates for Mechanical Turk studies are often assumed to be zero, but are usually higher [64]

- ☐ IP addresses, for identifying (reasonably rare) cases of multiple submissions and cases where responses may not have been independent (e.g. two respondents in the same room)
- ☐ Browser agent strings, which contain information about the platform and browser used to complete the study
- ☐ Demographics, because samples are not selected randomly from Mechanical Turk, citing demographics from previous studies on the platform may not be appropriate
- ☐ Whether participants discovered the experiment on a site outside of Mechanical Turk, and the URL where they found it, so participant discussions about the experiment can be monitored [51]

Importantly, sampling may occur differently between different marketplaces, depending on how the marketplaces are designed. MTurk makes tasks available to all eligible participants in reverse, but other services have sampling policies that prioritize workers with specific characteristics (e.g., experience) above and beyond what is requested by researchers—something that should be made as clear as possible by service providers.

4 Reproducibility

Science must be reproducible. Here we first discuss whether the results obtained from the crowdsourced convenience samples can be reproduced in other populations. Then we consider crowdsourcing in the light of the replication crisis.

4.1 Reproducible results.

Crowdsourced samples can be used along-side traditional university panel samples. That is, in addition to our "weird" (Western, educated, industrialized, rich and democratic) convenience samples of university students [65], we also have a second and quite different convenience sample of MTurkers (although they too are weird). Numerous studies have found comparable results when using these different samples. Classical experiments in judgment and decision making have been consistently replicated [3, 6, 8] and the psychometric properties of individual difference scales like the big five are usually excellent [7, 53]. Phenomena observed within many commonly used cognitive psychology paradigms are also observed within MTurk worker populations [35, 39, 40, 66].

More recent and ambitious efforts have examined the replicability of research findings using batteries of experimental studies administered to large samples of MTurk workers and other individuals have found similarly consistent results. In the many-labs project, the pattern of effects and null effects for 13 social psychology and decision making coefficients corresponded perfectly between concurrently collected student and MTurk samples [67]. Similarly

high replication rates have been observed within batteries of cognitive psychology experiments [39, 68]. More than 90% (33/36) of correlations between attitudes and personality measures did not differ across an MTurk sample and a representative sample collected for the American National Election Studies [69]. In political science, 82% (32 of 39) of effects and null effects observed in a high quality probability sample replicated in an MTurk sample, and the effect sizes observed across the two samples did not differ from each other in six of the seven remaining conditions [4] (see also a rate of 68% (25/37) in [70]).

4.2 Reproducibility and the replication crisis.

Many scientists are concerned that there is a "replication crisis" in science, and recently much effort has gone into fixing research practices [71]. While some researchers are, anecdotally, somewhat skeptical about online studies, we do not see crowdsourcing as a cause or fix for the replication crisis. Here we consider issues around speed of data collection and sample size.

It is unclear what effect the ease and speed of running studies on crowdsourcing platforms will have on reproducibility. On one hand, it could be argued that the ease with which studies can be run encourages more bad (that is, likely false) ideas to be tested, filling up the file drawer [72] while also increasing false positives. It could also be argued that the ease of starting and stopping data collection could influence decisions to continue or extend a study (p-hacking [73] but see [74], or take a Bayesian approach [75] but see [76]). On the other hand, lowering the time and resource costs of being wrong could make it easier for researchers—particularly those with limited access to participants or facing looming tenure deadlines—to let go of ideas that are less promising. Further, unlike traditional studies that might collect only a few participants per day, data peeking imposes real-time costs on MTurk for little gain in understanding how a study is unfolding. Whatever portion of data peeking that is motivated by excitement, rather than deliberate attempts at data manipulation, should simply go away if the sole benefit is a few hours advance notice about the studies likely outcome. The availability of crowdsourced samples cannot be a cause *per se* of questionable research practices—like any tool, it is the responsibility of the researcher to use it in a methodologically sound way.

The efficiency of crowdsourcing also allows larger samples—and we need larger sample sizes as much research is underpowered [77]. Under-powered studies are likely to unintentionally produce false positives [78, 79] and as sample sizes increase, results become less sensitive to some forms of researcher degrees of freedom [78]. Large samples are also important if the field is to move towards estimation of effect sizes [80], rather than just null hypothesis significance testing. For example, to estimate a medium effect size and exclude small and large effects from the confidence intervals, we need samples of at least 1,000 (http://datacolada.org/20). Crowdsourcing can deliver these sample sizes.

The availability of crowdsourced samples may also increase attempts to reproduce results observed by others. The speed and cost advantages of using crowds lower the investment required to replicate a study. The ability to recruit large samples also benefits replications, which typically require more participants than the study they seek to replicate [81]. Perhaps most importantly, a crowd provides a standardized common sample that all researchers can use. Failed replications are inherently ambiguous: either the finding in the original study

is"true" or the replication is"true" or the replication differed in a crucial but overlooked way [82–84]. As a result, failed replications inevitably lead to speculation about possibly overlooked differences between the original study and the replication attempt and often lead to follow up studies by either the original or replicating authors to rule out potential explanations. The ability to share exact materials and recruit samples drawn from the same population (see Box "Crowdsourcing Checklist for Authors, Reviewers, and Editors") reduces the number of potential differences between studies considerably (though, with the potential for complications, see A Tragedy of the Commons, below), simplifying this exercise for everyone involved.

5 Design and Ethics for Crowdsourcing

Running experiments on crowdsourcing platforms is not the same as running experiments in the laboratory with university participants. Crowdsourcing tasks tend to be of short duration (though there are exceptions [85]), participants are seldom taking part under exam conditions (with nearly one third being not alone, and one fifth reporting watching TV [63]) and there is almost no opportunity for participants to interact with the experimenter or ask questions, which means your experimental task needs to afford correct completion by design, without lengthy instructions.

There are ethical issues to consider too. Is "Click here to continue" good enough for informed consent? Do participants understand that they can withdraw from the study at any time? It is arguably easier just to close a browser window than to leave a laboratory with an experimenter present, but, anecdotally, workers differ in their beliefs about the consequences of abandoning a HIT. Do participants have a way to request the deletion of their data? This may be harder if they do not have a ready way to revisit old experiments. Remember WorkerIDs are not anonymous [86], so avoid posting them with data.

There have been calls for ethical rates of pay [55, 87]. The decision to pay workers more is primarily an ethical one. In general, data quality defined as providing reliable self-report and larger experimental effect sizes is usually not sensitive to compensation level [7, 39, 88], at least for American workers [89]. However, paying more will definitely increase the speed of data collection [3, 88] and may reduce participant attrition [39]. Payment may also induce workers to spend longer on tasks which require sustained effort, and will thus perform better when performance is correlated with effort [90, 91]. One potential downside of higher pay rates is that they may also attract the most keen and active participants, crowding out less experienced workers [51] and shrinking the population from which you are sampling [62].

6 A Tragedy of the Commons?

Amazon reports more than 500,000 registered workers on MTurk, but, like many online communities, the number of active workers at any given time is much smaller—at least one order of magnitude smaller [87]. The pool of workers available to a researcher is further limited, particularly when the sample drawn from the pool is small, because some workers are more likely to complete a study than other workers. A capture-recapture analysis using

data from seven laboratories across the world estimated that the average laboratory samples from a population of about 7,300 MTurk workers, with substantial overlap between the populations accessed by different laboratories [62]. This creates the very real possibility of exhausting the pool of available workers for a particular line of research.

Unlike traditional subject pools, there are no restrictions on how many surveys a worker may complete or how long they may remain in the pool. Many workers consider MTurk to be like a job, spending many hours per week completing surveys. The median MTurker reports participating in about 160 academic studies in the previous month [92], and many would complete more if they could [93]. Workers remain in the pool as long as they like, with about half of the workers replaced every seven months [62], but some remaining in the pool for years. The lack of restrictions on worker productivity and tenure leads to a small group of workers who are responsible for a large fraction of the observations obtained on MTurk research studies [63].

Researchers have also raised concerns about workers sharing knowledge about experiment with each other on forums or online discussion boards. About 60% of workers use forums and about 10% can directly contact at least one other MTurk worker [94]. To date, most available evidence suggests that this is rare. Perhaps 10% of workers are referred to HITs from sources outside of Mechanical Turk (mostly Reddit [51]). Within these communities there are strong norms about intentionally disclosing substantive information about a study. Although these norms are not always followed and people may not realize that certain information is substantive, by and large these violations are rare, with workers using these forums primarily to share instrumental features of tasks (e.g., well paying HITs or bad experiences with requesters) [61, 63]

Worker experience can be a problem because experience completing studies can influence behavior in future studies. Measures of performance will become inflated through practice effects. For example, the classic cognitive reflection test [95] is largely known to MTurk workers [96]. Consequently, worker experience is correlated with performance on this test, but not on logically equivalent novel questions [63, 97]. Observed effect sizes may decrease over time when one variable is correlated with prior experience, but another is not. For example, in series of experimental games conducted on MTurk, workers intuitive tendencies shifted from cooperation towards the optimal game-theoretic rational response [16, 98].

Familiarity with an experiment may also reduce effect sizes if information from other experimental conditions contaminates judgments. In a systematic replication of 12 studies, participating twice in the same two-condition experiment reduced effect sizes, particularly for those who were assigned to the alternative condition and when little time had elapsed between participations [99]. Moreover, there is at least some evidence of effects that only replicate with naïve participants [100]. Fortunately, experiments that examine many commonly used outcomes in cognitive psychology appear to be robust to prior exposure, with task performance improving over time, but between group differences persisting [68].

Participant non-naïveté may have effects on data quality that go beyond those of the task-specific. Familiarity with attention checks may make participants very good at dealing with any attention check (and better than University participants)—independent of familiarity with the specific check [101], which has led to the call for researchers to use novel attention checks, or to abandon them for this population altogether [47]. Participants can also gain familiarity with eligibility criteria used to prescreen for admission in one study

and use this information fraudulently to gain access to later studies with similar eligibility requirements [102]. Psychologists can contaminate the participant pool with deception manipulations which are particularly disliked by experimental economists [103]. Even in absence of outright deception, research procedures that try to disguise the true purpose of an experiment may be less effective with expert participants who are more aware of these procedures [104].

The small size of the pool of active participants and the prevalence of expert or professional participants on online labor markets can lead to a tragedy of the commons, where studies run in one laboratory can contaminate the pool for other laboratories running other studies. This possibility is compounded by the difficulty requesters face in communicating with each other, or even knowing what experiments other researchers have conducted.

Although there are real challenges in managing a shared pool, and researchers would certainly benefit from new tools to assist in this process, the problems it poses are manageable. Although asking workers whether they have completed similar experiments before is unlikely to be effective [99, 102], researchers can automatically prevent specific workers from completing related studies that they have posted by using qualifications [59]. In fact, TurkPrime can prevent any identified worker from completing a given experiment, making it possible for researchers to share lists of workers to avoid [1] (although it should be remembered that worker identification numbers can, in certain narrow circumstances, be linked to individuals and should thus be treated as personally identifying information [86]). Importantly, all of the available evidence suggests that, if anything, repeated exposure to research materials attenuates effects, which can be offset by increasing sample sizes and at worst leads to a larger proportion of false negatives. Whether repeated exposure to a research paradigm can produce false positives remains an empirical question.

Given the collective move by researchers from a large set of independent University-based samples to one shared MTurk sample, there is also the possibility that one big event, or a single decision by the crowd administrators, could affect the viability of data collection with the crowd. For example, in 2016, Amazon raised the MTurk commissions with short notice, increasing the price of conducting studies. Similarly, Amazon has made unpredictable changes in their policy, cutting off access for international requesters and workers. As we utilize MTurk more as a discipline, we have increased the systemic risk we all face—though diversifying across the newer platforms will reduce this risk.

7 Concluding Remarks

Crowdsourcing cognitive science offers us a new route for scientific advance in our field. We need to exploit the positive features of the platform whilst mitigating the weaknesses. We need to report carefully how we have used our chosen platform, and might find it particularly useful to pre-register aspects of our research. There will be innovations in the field as sensors increase beyond webcams to other internet of things devices (e.g., Fitbits and GPS) and locations outside the home and office. There will be innovations where people can contribute their machine recorded data like supermarket transactions or banking records in exchange for insight and control over their data. And we will need to understand how sharing a common pool of relatively expert participants influences our findings.

Author Note

Stewart has a close relative who works for Amazon, the owners of MTurk. This work was supported Economic and Social Research Council grants ES/K002201/1, ES/K004948/1, ES/N018192/1, and Leverhulme grant RP2012-V-022.

References

[1] Litman, L. *et al.* (2016) TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav. Res. Methods*

[2] Gosling, S. D. and Mason, W. (2015) Internet research in psychology. *Annu. Rev. Psychol.* 66, 877–902

[3] Berinsky, A. J. *et al.* (2012) Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Polit. Anal.* 20, 351–368

[4] Mullinix, K. J. *et al.* (2016) The generalizability of survey experiments. *J. Exp. Polit. Psychol.* 2, 109–138

[5] Kittur, A. *et al.* (2008) Crowdsourcing user studies with Mechanical Turk. In Czerwinski, M. *et al.*, eds., *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM

[6] Paolacci, G. *et al.* (2010) Running experiments on Amazon Mechanical Turk. *Judgm. Decis. Mak.* 5, 411–419

[7] Buhrmester, M. *et al.* (2011) Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives On Psychol. Sci.* 6, 3–5

[8] Horton, J. J. *et al.* (2011) The online laboratory: Conducting experiments in a real labor market. *Exp. Econ.* 14, 399–425

[9] Shank, D. B. (2016) Using crowdsourcing websites for sociological research: The case of Amazon Mechanical Turk. *Am. Sociol.* 47, 47–55

[10] Shapiro, D. N. *et al.* (2013) Using Mechanical Turk to study clinical populations. *Clinical Psychol. Sci.* 1, 213–220

[11] Goodman, J. K. and Paolacci, G. (2017) Crowdsourcing consumer research. *J. Consum. Res.*

[12] Bentley, J. W. (2017) Challenges with Amazon Mechanical Turk research in accounting

[13] Stritch, J. M. *et al.* (2017) The opportunities and limitations of using Mechanical Turk (Mturk) in public administration and management scholarship. *Int. Public. Manag. J.*

[14] Scott, K. and Schulz, L. (2017) Lookit (part 1): A new online platform for developmental research. *Open Mind* 1, 4–14

[15] Tran, M. *et al.* (2017) Online recruitment and testing of infants with Mechanical Turk. *J. Exp. Child Psychol.* 156, 168–178

[16] Rand, D. G. *et al.* (2014) Social heuristics shape intuitive cooperation. *Nature Commun.* 5, e3677

[17] Arechar, A. *et al.* (2017) Conducting interactive experiments online. *Exp. Econ.*

[18] Balietti, S. (2016) nodeGame: Real-time, synchronous, online experiments in the browser. *Behav. Res. Methods*

[19] Yu, L. and Nickerson, J. V. (2011) Cooks or cobblers? crowd creativity through combination. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1393–1402. ACM

[20] Kim, J. *et al.* (2016) Mechanical novel: Crowdsourcing complex work through reflection and revision. *Comput. Res. Repository*

[21] Morris, R. R. and Picard, R. (2014) Crowd-powered positive psychological interventions. *J. Positive Psychol.* 9, 509–516

[22] Bigham, J. P. *et al.* (2010) VizWiz: Nearly real-time answers to visual questions. In Perlin, K. *et al.*, eds., *Proceedings of the 23nd annual ACM symposium on user interface software and technology*, pages 333–342. ACM, New York

[23] Meier, A. *et al.* (2011) Usability of residential thermostats: Preliminary investigations. *Build. Environ.* 46, 1891–1898

[24] Boynton, M. H. and Richman, L. S. (2014) An online diary study of alcohol use using Amazon's Mechanical Turk. *Drug Alcohol Rev.* 33, 456–461

[25] Dorrian, J. *et al.* (2017) Morningness/eveningness and the synchrony effect for spatial attention. *Accident Anal. Prev.* 99, 401–405

[26] Benoit, K. *et al.* (2016) Crowd-sourced text analysis: Reproducible and agile production of political data. *Am. Polit. Sci. Rev.* 110, 278–295

[27] Mueller, P. and Chandler, J. (2012) Emailing workers using Python

[28] Reimers, S. and Stewart, N. (2015) Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behav. Res. Methods* 47, 309–327

[29] Reimers, S. and Stewart, N. (2016) Auditory presentation and synchronization in Adobe Flash and HTML5/JavaScript Web experiments. *Behav. Res. Methods* 48, 897–908

[30] de Leeuw, J. (2015) jspsych: A javascript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* 47, 1–12

[31] Gureckis, T. M. *et al.* (2016) psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behav. Res. Methods* 48, 829–842

[32] Stoet, G. (2010) PsyToolkit: A software package for programming psychological experiments using Linux. *Behav. Res. Methods* 42, 1096–1104

[33] Stoet, G. (2017) Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teach. Psychol.* 44, 24–31

[34] Schubert, T. W. *et al.* (2013) ScriptingRT: A software library for collecting response latencies in online studies of cognition. *PLoS One* 8

[35] de Leeuw, J. R. and Motz, B. A. (2016) Psychophysics in a web browser? Comparing response times collected with javascript and psychophysics toolbox in a visual search task. *Behav. Res. Methods* 48, 1–12

[36] Neath, I. *et al.* (2011) Response time accuracy in apple macintosh computers. *Behav. Res. Methods* 43, 353–362

[37] Ulrich, R. and Giray, M. (1989) Time resolution of clocks: Effects on reaction time measurement-Good news for bad clocks. *Brit. J. Math. Stat. Psy.* 42, 1–12

[38] Brand, A. and Bradley, M. T. (2012) Assessing the effects of technical variance on the statistical outcomes of web experiments measuring response times. *Soc. Sci. Comput. Rev.* 30, 350–357

[39] Crump, M. J. *et al.* (2013) Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8, e57410

[40] Hilbig, B. E. (2016) Reaction time effects in lab- versus web-based research: Experimental evidence. *Behav. Res. Methods* 48, 1718–1724

[41] Semmelmann, K. and Weigelt, S. (2016) Online psychophysics: Reaction time effects in cognitive experiments. *Behav. Res. Methods*

[42] Slote, J. and Strand, J. F. (2016) Conducting spoken word recognition research online: Validation and a new timing method. *Behav. Res. Methods* 48, 553–566

[43] Lutz, J. (2016) The validity of crowdsourcing data in studying anger and aggressive behavior a comparison of online and laboratory data. *Soc. Psychol.* 47, 38–51

[44] Majima, Y. *et al.* (2017) Conducting online behavioral research using crowdsourcing services in Japan. *Front. Psychol.* 8, 378

[45] Peer, E. *et al.* (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* 46, 1023–1031

[46] Crone, D. L. and Williams, L. A. (2016) Crowdsourcing participants for psychological research in Australia: A test of Microworkers. *Aust. J. Psychol.*

[47] Peer, E. *et al.* (2017) Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Soc. Psychol.* 70, 153–163

[48] Estellés-Arolas, E. and González-Ladrzón-De-Guevara, F. (2012) Towards an integrated crowdsourcing definition. *J. Inf. Sci.* 38, 189–200

[49] Sulser, F. *et al.* (2014) Crowd-based semantic event detection and video annotation for sports videos. In Redi, J. and Lux, M., eds., *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 63–68. ACM, New York

[50] Casler, K. *et al.* (2013) Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav.* 29, 2156–2160

[51] Casey, L. *et al.* (2017) Intertemporal differences among mturk worker demographics

[52] Levay, K. E. *et al.* (2016) The demographic and political composition of Mechanical Turk samples. *Sage Open*

[53] Behrend, T. S. *et al.* (2011) The viability of crowdsourcing for survey research. *Behav. Res. Methods* 43, 800–813

[54] Arditte, K. A. *et al.* (2016) The importance of assessing clinical phenomena in Mechanical Turk research. *Psychol. Assessment* 28, 684

[55] Goodman, J. K. *et al.* (2013) Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *J. Behav. Decis. Making.* 26, 213–224

[56] Kosara, R. and Ziemkiewicz, C. (2010) Do Mechanical Turks dream of square pie charts? In Sedlmair, M. *et al.*, eds., *Proceedings of the 3rd BELIV'10 Workshop Beyond Time and Errors: Novel Evaluation Methods for Information Visualisation*, pages 63–70. ACM, New York

[57] Johnson, D. R. and Borden, L. A. (2012) Participants at your fingertips: Using Amazon's Mechanical Turk to increase student-faculty collaborative research. *Teach. Psychol.* 39, 245–251

[58] Veilleux, J. C. *et al.* (2014) Negative affect intensity influences drinking to cope through facets of emotion dysregulation. *Pers. Indiv. Differ.* 59, 96–101

[59] Chandler, J. and Shapiro, D. (2016) Conducting clinical research using crowdsourced convenience samples. *Annu. Rev. Clin. Psycho.* 12, 53–81

[60] Arechar, A. A. *et al.* (2016) Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk

[61] Wang, X. *et al.* (2017) A community rather than a union: Understanding self-organization phenomenon on Mturk and how it impacts Turkers and requesters. In *Association for Computing Machinery CHI'17 Conference*, pages 2210–2216. ACM, New York

[62] Stewart, N. *et al.* (2015) The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgm. Decis. Mak.* 10, 479–491

[63] Chandler, J. *et al.* (2014) Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behav. Res. Methods* 46, 112–130

[64] Zhou, H. and Fishbach, A. (2016) The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *J. Pers. Soc. Psychol.*

[65] Henrich, J. *et al.* (2010) Most people are not WEIRD. *Nature* 466, 29–29

[66] Simcox, T. and Fiez, J. A. (2014) Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behav. Res. Methods* 46, 95–111

[67] Klein, R. A. *et al.* (2014) Investigating variation in replicability: A "many labs" replication project. *Soc. Psychol.* 45, 142–152

[68] Zwann, R. A. *et al.* (2017) Some psychological effects replicate even under potentially adverse conditions

[69] Clifford, S. *et al.* (2015) Are samples drawn from Mechanical Turk valid for research on political ideology? *Res. Polit.* 2

[70] Coppock, A. (2016) Generalizing from survey experiments conducted on Mechanical Turk: A replication approach

[71] Munafo, M. R. *et al.* (2017) A manifesto for reproducible science. *Nature Hum. Behav.* 1, 0021

[72] Rosenthal, R. (1979) The file drawer problem and tolerance for null results. *Psychol. Bull.* 86, 638–641

[73] Simmons, J. P. *et al.* (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366

[74] Frick, R. W. (1998) A better stopping rule for conventional statistical tests. *Behav. Res. Methods, Instruments, & Computers* 30, 690–697

[75] Kruschke, J. K. (2011) *Doing Bayesian Data Analysis: A Tutorial with R and BUGS.* Academic Press, Burlington, MA

[76] Simonsohn, U. (2014) Posterior-hacking: Selective reporting invalidates Bayesian results also

[77] Cohen, J. (1988) *Statistical power analysis for the behavioral sciences.* Erlbaum, Hillsdale, NJ, 2nd ed.

[78] Button, K. S. *et al.* (2013) Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* 14, 365–376

[79] Open Sci Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349

[80] Cumming, G. (2014) The new statistics: Why and how. *Psychol. Sci.* 25, 7–29

[81] Simonsohn, U. (2015) Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* 26, 559–569

[82] Open Sci Collaboration (2012) An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. On Psychol. Sci.* 7, 657–660

[83] Schwarz, N. and Strack, F. (2014) Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Soc. Psychol.* 45, 305–306

[84] Stroebe, W. and Strack, F. (2014) The alleged crisis and the illusion of exact replication. *Perspect. On Psychol. Sci.* 9, 59–71

[85] Mor, S. *et al.* (2013) Identifying and training adaptive cross-cultural management skills: The crucial role of cultural metacognition. *Acad. Manag. Learn. Edu.* 12, 139–161

[86] Lease, M. *et al.* (2013) Mechanical Turk is not anonymous

[87] Fort, K. *et al.* (2011) Amazon Mechanical Turk: Gold mine or coal mine? *Comput. Linguist.* 37, 413–420

[88] Mason, W. and Watts, D. J. (2010) Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter* 11, 100–108

[89] Litman, L. *et al.* (2015) The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behav. Res. Methods* 47, 519–528

[90] Aker, A. *et al.* (2012) Assessing crowdsourcing quality through objective tasks. In Chair), N. C. C. *et al.*, eds., *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA)

[91] Ho, C.-J. *et al.* (15) Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, pages 419–429. International World Wide Web Conferences Steering Committee

[92] Kees, J. *et al.* (2017) An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *J. Advertising* 46, 141–155

[93] Berg, J. (2016) Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers. *Comp. Labor Law & Pol. J.* 37

[94] Yin, M. *et al.* (2016) The communication network within the crowd. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1293–1303. International World Wide Web Conferences Steering Committee

[95] Frederick, S. (2005) Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42

[96] Thompson, K. S. and Oppenheimer, D. M. (2016) Investigating an alternate form of the cognitive reflection test. *Judgm. Decis. Mak.* 11, 99–113

[97] Finucane, M. L. and Gullion, C. M. (2010) Developing a tool for measuring the decision-making competence of older adults. *Psychol. Aging* 25, 271–288

[98] Mason, W. *et al.* (2014) Long-run learning in games of cooperation. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, pages 821–838. ACM, New York

[99] Chandler, J. *et al.* (2015) Using non-naïve participants can reduce effect sizes. *Psychol. Sci.* 26, 1131–1139

[100] DeVoe, S. E. and House, J. (2016) Replications with MTurkers who are naïve versus experienced with academic studies: A comment on Connors, Khamitov, Moroz, Campbell, and Henderson (2015). *Journal of Experimental Soc. Psychol.* 67, 65–67

[101] Hauser, D. J. and Schwarz, N. (2015) Attentive turkers: Mturk participants perform better on online attention checks than subject pool participants. *Behav. Res. Methods*

[102] Chandler, J. and Paolacci, G. (2017) Lie for a dime: When most prescreening responses are honest but most study participants are imposters. *Soc. Psychol.*

[103] Hertwig, R. and Ortmann, A. (2001) Experimental practices in economics: A methodological challenge for psychologists? *Behav. Brain. Sci.* 24, 383–451

[104] Krupnikov, Y. and Levin, A. S. (2014) Cross-sample comparisons and external validity. *J. Exp. Polit. Psychol.* 1, 59–80