Published in final edited form as:

Addiction. 2020 October ; 115(10): 1960–1968. doi:10.1111/add.15032.

Mechanical Turk Data Collection in Addiction Research: Utility, Concerns and Best Practices

Alexandra M. Mellis^a, Warren K. Bickel^a

^aAddiction Recovery Research Center, Fralin Biomedical Research Institute at VTC, VA, USA

Abstract

Aims: Amazon Mechanical Turk (MTurk) provides a crowdsourcing platform for the engagement of potential research participants with data collection instruments. This review (1) provides an introduction to the mechanics and validity of MTurk research; (2) gives examples of MTurk research; and (3) discusses current limitations and best practices in MTurk research.

Methods: We review four use cases of MTurk for research relevant to addictions: 1) the development of novel measures, 2) testing interventions, 3) the collection of longitudinal use data to determine the feasibility of longer-term studies of substance use, and 4) the completion of large batteries of assessments to characterize the relationships between measured constructs. We review concerns with the platform, ways of mitigating these, and important information to include when presenting findings.

Results: MTurk has proven to be a useful source of data for behavioral science more broadly, with specific applications to addiction science. However, it is still not appropriate for all use cases, such as population-level inference. To live up to the potential of highly transparent, reproducible science from MTurk, researchers should clearly report inclusion/exclusion criteria, data quality checks and reasons for excluding collected data, how and when data was collected, and both targeted and actual participant compensation.

Conclusions: Although online survey research is not a substitute for random sampling or clinical recruitment, the MTurk community of both participants and researchers has developed multiple tools to promote data quality, fairness, and rigor. Overall, MTurk has provided a useful source of convenience samples despite its limitations and has demonstrated utility in the engagement of relevant groups for addiction science.

Keywords

Amazon Mechanical Turk; addiction; crowdsourcing; survey research; online

Correspondence concerning this article should be addressed to Warren K. Bickel, Fralin Biomedical Research Institute at VTC, Roanoke, VA, 24016, wkbickel@vtc.vt.edu.

Declaration of interests: none

Introduction

Since its launch in 2005, Amazon Mechanical Turk (MTurk) has provided for a source for both data and debate in the behavioral sciences. Boasting a user base of hundreds of thousands of individuals, MTurk has fueled a spate of survey-based research in the social sciences. Critically, however, concerns with the nature of crowdsourcing websites, the quality of data collection, and the particulars of the MTurk labor market remain. Much ink has been spilled reviewing the use of MTurk in academic research (1–3), including in-depth review with specific applications to addiction science (4). Here, we seek to provide a brief overview of research on MTurk, outline four use cases in which MTurk has usefully addressed research questions, and make best practice recommendations for readers as well as addressing pitfalls in research conducted on Mturk.

The Mechanics of the Mechanical Turk

Amazon Mechanical Turk serves as a clearinghouse for entities who need tasks completed and individual workers offering their labor, in a system known as crowdsourcing. Requesters, such as researchers, can post data collection instruments as Human Intelligence Tasks, or HITs. These HITs typically take the form of a link from the MTurk website to a survey or other data collection service, such as Qualtrics. A labor pool of available Workers, or potential research participants, can then complete these HITs for compensation. These payments are typically delivered as both a base payment and a bonus, starting at \$0.01 for task completion. Requesters can select from available workers through both native and Requester-generated restrictions on access to data collection instruments, called qualifications. Requesters may, for example, select a qualification that only Workers with a past approval rating above a certain percent (indicating that over all of the Worker's previous HITs, that percent were accepted and therefore received base compensation) may complete their HIT.

Despite being initially developed to support computing projects (for example, by identifying and removing duplicate product listings on online retail marketplaces), MTurk was quickly identified for its potential contribution to research (see Figure 1). Thousands of manuscripts have been published since the website's launch. A seminal series of reviews (2) drew attention to MTurk for its ability to provide data of comparable reliability to data collected in-person. Indeed, despite the existence of multiple alternatives to MTurk (including comparable crowdsourcing websites such as CrowdFlower, and research-focused participant pools such as Prolific Academic (5); for a more complete review of online crowdsourcing alternatives to MTurk, see (6)), MTurk remains the most highly used platform in the behavioral and social sciences. However, given the ease of running behavioral research projects on MTurk and similar platforms, concerns are growing about the possibility of low-quality studies proliferating. These concerns about the validity of data collected from MTurk are typically shared more generally with survey-based research performed in convenience samples and may be mitigated with careful research design. Furthermore, new research implemented on the platform (see "Four Use Cases for MTurk") have supported the use of MTurk for use in research that would otherwise not be feasible for in-person data collection, demonstrating the rewards of these careful designs.

Comparability between MTurk and Other Convenience Samples

Multiple studies have sought to replicate a battery of psychological assessments in both lab-based and online samples, including samples of unpaid volunteers (7) and in comparisons between computer-based tasks completed on MTurk and purely behavioral tasks (8). One study systematically compared performance across 10 tasks (including Stroop, flanker, and multiple learning tasks) among MTurk workers to published results from in-person studies. Overall, task performance was comparable in the online sample as in previously published results, especially after the introduction of instruction comprehension checks (9). Another study compared performance on two decision-making tasks, the Asian disease problem and the physician problem, and found similar biases in decision-making among individuals recruited from MTurk, alternative crowdsourcing platforms (including CrowdFlower and ProlificAcademic), and a midwestern university (10). Direct comparisons have also been made between data collected on MTurk and data collected via online panels (11). We highlight here one exemplar study that replicated previously-observed effects, relevant to addiction, in an MTurk sample.

Delay discounting, or the phenomenon of devaluing reinforcers as a function of delay to their receipt, has been well documented in both nonhuman and human research. Specific findings related to discounting rates and addiction, including observation of steeper delay discounting among individuals with substance use disorders compared to those without, have also been replicated and demonstrated in meta-analyses (12,13). Johnson and colleagues (14) attempted to replicate six phenomena previously observed in delay discounting research. Five of the six phenomena, including observation of hyperbolic discounting rates, negative correlation between discounting rate and age and education, and steeper discounting of consumable reinforcers (cigarettes) than of money. After controlling for education, this study did not observe statistically significant differences in discount rate between substance users (here, smokers) and controls. However, this was attributed to methodological choices in the survey design (included a longest possible delay to reward receipt of 24 hours), rather than differences in the MTurk population.

Overall, a review of 35 manuscripts comparing performance between MTurk and other samples identified that responses are comparable across populations (15). Some additional research has specifically explored differences in survey completion strategy between MTurk and other samples, revealing possibly higher rates of survey satisficing among online, for-pay survey completers (16), and sensitivity of MTurk participants to slight changes in instructions (17). However, by and large, a replication-based approach has indicated that results observed on MTurk can generalize to other laboratory populations (18).

Concerns and Limitations in Mechanical Turk Research

We note that although MTurk has its place for recruitment of convenience samples, including in addiction research, several cautions must be mentioned. Specifically, recent concerns have been articulated about rates of low-quality data and nonsensical, distracted, or fraudulent responding, which may be increasing since 2018 (19). Similarly, online studies may allow recruitment of individuals who do not meet study eligibility criteria (e.g.,

participants outside of the US for studies intended to be US-based). Indeed, rates of Virtual Private Server (VPS) use may have recently increased on MTurk (20). These systems allow users to “spoof” IP addresses to, for example, appear to respond as if located in the United States, and can be addressed using standard survey software such as Qualtrics (21). Another, potentially graver concern is that although the MTurk platform itself describes a potential pool of 500,000 participants, the effective sample may be much smaller. One long-term study of the demographics of the MTurk workforce used capture-recapture methods to monitor changes in MTurk population characteristics and size between 2015 and 2018. Critically, although substantial numbers of workers are observed repeating multiple studies, this analysis estimated that tens of thousands of new workers arrive on the platform each year, and the average worker half-life is around 400 days. However, at any given time the estimated population available is under 2,500 workers (22). Alternative analyses have placed this number at around 7,300 (23). If the overall population of available workers is small, experimental history must be considered in recruitment and analysis. As MTurk is increasingly used to study the effects of interventions, the possibility of non-naivete must also be considered in the blinding of participants to active and control groups, or must be accounted for analytically (as in 24, which explicitly interrogated participant’s expectation of being assigned to either an active or a control group). Note, however, that the effective pool of available workers may be increased through multiple methodological choices, such as posting HITs in larger batches (making a larger number of tasks available at once and disallowing repeated submissions), excluding past participants through assigned Qualifications, and allowing Workers who have completed fewer HITs total to participate.

Finally, any researcher must remember that MTurk is fundamentally a source for convenience sampling, and does not offer a representative sample of either whole populations or of particular clinical groups. The MTurk population has been previously shown to be younger and less likely to be fully employed than national averages (23) and to have different financial attitudes than those from other community samples (25). Within the US, MTurk workers are also more likely to be politically liberal (26). Due to the nature of the platform, workers are also typically computer-savvy, an effect which is exacerbated among the most prolific Workers by the need for browser extensions and supportive scripts to help workers reduce downtime between HITs and maximize earnings. However, MTurk has potential (as does other online, social-media-based recruitment (27)) to offer a nonprobability-based method for oversampling participants with particular conditions relevant to addiction, such as heavy drinkers or individuals reporting chronic pain—similar to flyer-based community recruitment of individuals with a history of substance use. These sampling methods must be carefully considered to recruit ideal participants (e.g., recruiting participants based on drinking rate, rather than AUDIT score, for an intervention targeting drinking rate itself (28)). Individuals may also be recruited to complete extensive screening batteries to allow for the assignment of researcher-defined Qualifications. Although online recruitment precludes biological verification of responses, these assessments may be appropriate for any inclusion criterion typically assessed using self-report screeners. Critically, no matter how carefully convenience sampling methods are employed, these methods should not be used to make population-level inferences.

Four Use Cases for MTurk

Validating an Instrument.

Emphasizing the ability of online research to reach large sample sizes at low cost, MTurk has proven valuable in the development of multiple instruments in psychometric research. In one study, 647 participants were screened from MTurk to identify 300 self-identifying chronic pain patients to develop a measure of the likelihood of taking a novel analgesic based on varying levels of pain relief and the probability of addiction. The external validity of this measure was demonstrated, identifying that individuals at the highest risk for opioid misuse using a standard clinical measure were also the most likely to take a novel analgesic with high potential for addiction (29).

Testing an Intervention.

One additional use for MTurk is in the development and validation of interventions, rather than instruments. Narrative interventions for addiction (30) may be particularly well-suited to development via crowdsourcing. One narrative intervention, episodic future thinking, requires participants to engage in self-relevant, positive prospection to ameliorate steep delay discounting rates. Episodic future thinking has been shown to reduce discounting and valuation of alcohol (31). Additionally, compared to a control condition, episodic future thinking reduced delay discounting and self-administration of cigarettes (32); see Figure 2, top panel. In an extension of this work, Stein and colleagues (24) examined the effects of episodic future thinking in an MTurk sample of 117 cigarette smokers, all of whom passed internal consistency measures serving as data quality checks. Participants were randomly assigned to either an episodic future thinking or control condition and completed assessments of delay discounting as well as a hypothetical cigarette purchase task and craving assessment. Participants in the episodic future thinking group demonstrated reduced discounting and greater intensity of demand for cigarettes; see Figure 2, bottom panel. This study demonstrates the use of crowdsourcing particularly to test whether a theoretically-informed intervention alters hypothesized process measures.

Not all studies developing interventions for addiction have found success through MTurk. One study examined the feasibility of using MTurk to recruit for an online intervention on drinking behaviors. In approximately 3 hours, 1,252 participants had screened for the initial HIT. Three months after screening, 423 eligible participants who had consented to additional randomization were prompted to engage with either an online brief intervention or a control condition, and 360 participants completed a follow-up survey. Fewer than 40% of participants in the intervention group engaged with a brief intervention (28). Critically, engagement with this brief drinking intervention was not compensated, unlike engagement with episodic future thinking above. MTurk workers may be particularly sensitive to time constraints, and unlikely to engage with interventions unless directly compensated for their time engaged in research. Researchers must consider this when developing MTurk studies (see Review of Best Practices).

Collecting Longitudinal Data.

Another study demonstrated the feasibility of MTurk for the collection of longitudinal data concerning alcohol use (33). A sample of 278 adult drinkers was recruited to participate in 18 weekly assessments of alcohol and soda use. Compensation was offered both as payment for survey completion and as a probabilistic reward through a lottery for additional bonuses. Overall, the average response rate over the 18-week period was 73%, and 43% of participants completed all 18 weekly assessments. Participants reported these weekly assessments as overall enjoyable, convenient, fairly compensated, and easy to complete. The validity of this data was supported both by replication of previously-published relationships with drinking reports and demographic variables observed in conventional lab samples; and in reporting of higher drinking on weekends. Taken together, these results support the use of MTurk to recruit populations for repeated engagement with compensated assessment batteries.

Completing Intensive Batteries.

Finally, a large-scale examination of the reliability of self-control measures has provided a “stress test” for the engagement of participants through MTurk. In an ambitious research project examining underlying features defining the construct of self-control, participants have been recruited from MTurk to engage in lengthy batteries of self-control measures (34). Participants were given a week to complete a 10-hour battery of 63 assessments, including both task-based and survey-based measures of self-control, in addition to demographic variables assessing health behaviors. In the first presentation of this battery, 84% of participants completed all measures and were compensated \$65-\$75. Subsequently, 242 participants were invited to re-complete this battery after 2–4 months (35), and 150 participants completed these assessments. Test-retest reliability of self-control measures in this follow-up study was high, suggesting both stability in the construct of self-control and in the validity of data collected through MTurk.

Summary of Appropriate Use Cases.

Overall, MTurk (and other crowdsourcing clearinghouses) provide useful tools for researchers, which must be used thoughtfully. The four use cases described above demonstrate examples when MTurk provided additional insight to the addiction science community, but are not exhaustively representative of the types of research suited to the platform. Nor, however, do they indicate that all similar work would be appropriate. MTurk is especially well-suited to:

- developing measures (as in use case 1), testing specific hypotheses of theories (use case 2),
- testing generalizability of effects to online or self-administered platforms (use case 2), recruiting participants for self-report assessments (use case 3),
- and recruiting large samples for whom in-person data collection may not be feasible for all research groups (use case 4).

In general, these are circumstances where in-person data collection in non-clinical samples would be appropriate, but not necessarily feasible. However, MTurk is *not* well-suited to the

generation of population-level estimates (e.g., of the frequency of co-use of particular substances), collection of data where falsified reports may appear incentivized (e.g., recruitment of participants for contingency management interventions, without alternative verification of self-reported use), and paradigms in which the data collection environment may challenge the conclusions that may be drawn from research (e.g., where varying levels of distraction may hamper interpretation).

Review of Best Practices

A key consideration in best practices for research must be the ethical treatment of research participants. Although MTurk has grown in popularity largely as a source of “inexpensive, high-quality data” (2), participants are still volunteering for research and sharing information with minimal personal benefit. Participants must be fairly compensated, and recent manuscripts have demonstrated model behaviors in reporting both planned and actual hourly compensation rates (33). Fair pay is a central ethical concern for this research, as is the ability of participants to withdraw without prejudice (36). Median pay across MTurk is approximately \$2/hr (37) after accounting for the time required for Workers to identify HITs. Compensation should be thoughtfully reviewed during study design and IRB approvals, reported in manuscripts, and is increasingly being made transparent to workers, by workers. Researchers can engage with the MTurk worker community through multiple community websites and tools, including TurkOpticon/Dynamo, TurkerNation, TurkerView, and TurkerHub. Just as researchers may develop qualifying (or disqualifying) criteria for particular workers based on other Requester’s feedback (e.g., “<95% of HITs approved historically”), workers have developed systems to rate requesters. Workers have reported fairness in approval/rejection of work, prompt compensation, minimum payments, and responsiveness of requesters as primary concerns (38). At times, researcher expectations of experimental control are at odds with worker expectation of compensation (e.g., when the ideal inclusion criteria would best be assessed through a 20-question screener, but participants expect to be compensated for screeners longer than a few questions). Given the overall low burden of collecting data from MTurk, researchers should strive to consider the norms and preferences of the microtask, crowdsourcing economy when designing experimental and compensation schemes. These expectations, and challenges faced especially when working with institutional Requesters, have been articulated by Worker communities themselves (Brian McInnis Cornell University, Ith...)

Additionally, multiple best practices have been developed to promote data quality in MTurk studies from a research perspective. Posting research instruments as single, large batches of HITs may increase the effective population size available, reducing the probability of “repeat participants” in research where naivete is an important consideration (23). By posting larger batches and disallowing repeated submissions to the same batch, researchers can more easily ensure that each participant may only respond once, leading “deeper” sampling of the Worker population. These include both the use of checks to ensure that participants understand task instructions clearly (9), and use of an inclusion criterion based on Worker’s “reputation,” or total percentage of approved HITs (39), in addition to other attention check questions (40) and task-specific data quality assessments (e.g., 41,42). Expecting a clear presentation of inclusion/exclusion criteria, and the number of collected datasets excluded

for each reason will increase rigor. Inclusion criteria must also be plainly described to ensure accurate reporting of behavior patterns (e.g., “self-reported heavy drinking in the past week”) compared to diagnoses (e.g., “alcohol use disorder”), which typically cannot be verified in crowdsourced samples. Additionally, manuscripts reporting MTurk studies have begun to present specific dates and times of data collection, to ensure that researchers clearly articulate cases where additional participants were added post hoc (e.g., to increase sample size and observe a smaller-than-expected effect). These expectations may help improve replicability and reduce the risk of false-positive reports.

The potential for highly transparent, highly reproducible science.

As new tools are developed for use within the MTurk environment, additional methods may emerge for researchers to help each other manage the “crowd” component of crowdsourcing research. Previous efforts have combined data collected from MTurk across multiple labs to determine the possible effects of experimental history (23). Analysis of open-source data has also been critical to identifying common patterns used among participants employing VPSs to circumvent location restrictions, and helped to address scientific questions of both theoretical and methodological import (35). Taken together, the entirely-online MTurk environment means that both individual researchers and broader scientific efforts benefit from steps taken towards transparency both in data and in analyses. The MTurk Application Programming Interface (API) permits scripted experiment presentation, review of responses, and participant contact/compensation. These command-line tools allow for scripting of procedures such as posting of HITs, assessing and granting Qualifications, approving or rejecting responses, and granting bonuses, in multiple languages. This could theoretically allow for fully open source research, in which the complete code for both experimental procedures and analysis are made available for replication (as in (25)). We note, however, that the MTurk landscape is constantly changing--as of this review, the package for MTurk study scripting in the R language (47) has been deprecated and replaced by a new version which relies on Amazon’s Python tools, pyMturkR (48). As new methods and norms in crowdsourcing platforms emerge (e.g., the rating of requesters based on pay, rejection rates, and responsiveness, as on Turk Nation) and competitors innovate with additional tools to support research (e.g., by screening for naivete to experimental manipulations across tasks, as Prolific Academic proposes to do), scientists will both enjoy additional tools for research and additional responsibilities to use them wisely.

Conclusion

Crowdsourcing platforms like MTurk provide a useful tool for convenience sampling in behavioral research. The platform permits deployment of both task-based and survey-based data collection instruments and has been shown to successfully engage special subpopulations of participants relevant for addiction research (e.g., regular drinkers and those who self-report chronic pain). The limitations of this platform tend to be shared with other convenience samples--namely, non-random opt-in responding--but special attention must be paid to the limits inherent to online and remote assessment. MTurk is best understood as a system for targeted sampling of these special populations, rather than as a

source for clinical populations (requiring verification through personal health information) or representative sources of population estimates.

Acknowledgments

Funding: Preparation of this manuscript was supported by the National Institutes of Health, through the National Institute on Drug Abuse (R01 DA034755) and by the Science of Behavior Change Common Fund Program through an award administered by the National Institute of Diabetes and Digestive and Kidney Diseases (1UH2DK109543).

References

1. Rand DG. The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *J Theor Biol.* 2012 4 21;299:172–9. [PubMed: 21402081]
2. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspect Psychol Sci.* 2011 1;6(1):3–5. [PubMed: 26162106]
3. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk [Internet]. Vol. 44, *Behavior Research Methods.* 2012 p. 1–23. Available from: 10.3758/s13428-011-0124-6 [PubMed: 21717266]
4. Strickland JC, Stoops WW. The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Exp Clin Psychopharmacol.* 2019 2;27(1):1–18. [PubMed: 30489114]
5. Palan S, Schitter C. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance.* 2018 3 1;17:22–7.
6. Peer E, Brandimarte L, Samat S, Acquisti A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *J Exp Soc Psychol.* 2017 5 1;70:153–63.
7. Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, Wilmer JB. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychon Bull Rev.* 2012 10;19(5):847–57. [PubMed: 22829343]
8. Casler K, Bickel L, Hackett E. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput Human Behav.* 2013;29(6):2156–60.
9. Crump MJC, McDonnell JV, Gureckis TM. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One.* 2013 3 13;8(3):e57410. [PubMed: 23516406]
10. Paolacci G, Chandler J, Ipeirotis PG. Running Experiments on Amazon Mechanical Turk. 2010 6 24 [cited 2018 Mar 1]; Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1626226
11. Chandler J, Rosenzweig C, Moss AJ, Robinson J, Litman L. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behav Res Methods [Internet].* 2019 9 11; Available from: 10.3758/s13428-019-01273-7
12. MacKillop J, Amlung MT, Few LR, Ray LA, Sweet LH, Munafò MR. Delayed reward discounting and addictive behavior: a meta-analysis. *Psychopharmacology .* 2011 8;216(3):305–21. [PubMed: 21373791]
13. Amlung M, Vedelago L, Acker J, Balodis I, MacKillop J. Steep Delay Discounting and Addictive Behavior: A Meta-Analysis of Continuous Associations. *Addiction [Internet].* 2016 7 23; Available from: 10.1111/add.13535
14. Johnson PS, Herrmann ES, Johnson MW. Opportunity costs of reward delays and the discounting of hypothetical money and cigarettes. *J Exp Anal Behav.* 2015 1;103(1):87–107. [PubMed: 25388973]
15. Mortensen K, Hughes TL. Comparing Amazon's Mechanical Turk Platform to Conventional Data Collection Methods in the Health and Medical Research Literature. *J Gen Intern Med.* 2018 4;33(4):533–8. [PubMed: 29302882]
16. Hamby T, Taylor W. Survey Satisficing Inflates Reliability and Validity Measures: An Experimental Comparison of College and Amazon Mechanical Turk Samples. *Educ Psychol Meas.* 2016 12;76(6):912–32. [PubMed: 29795893]

17. Hauser DJ, Schwarz N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav Res Methods*. 2016 3;48(1):400–7. [PubMed: 25761395]
18. Coppock A, McClellan OA. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*. 2019 1 1;6(1):2053168018822174.
19. Chmielewski M, Kucker SC. An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Soc Psychol Personal Sci*. 2019 10 10;1948550619875149.
20. Kennedy R, Clifford S, Burleigh T, Waggoner P. The shape of and solutions to the MTurk quality crisis. Available at [Internet]. 2018; Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3272468
21. Winter N, Burleigh T, Kennedy R, Clifford S. A Simplified Protocol to Screen Out VPS and International Respondents Using Qualtrics [Internet]. 2019 [cited 2019 May 29]. Available from: <https://papers.ssrn.com/abstract=3327274>
22. Difallah D, Filatova E, Ipeirotis P. Demographics and Dynamics of Mechanical Turk Workers. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM; 2018 p. 135–43. (WSDM '18).
23. Stewart N, Ungemach C, Harris AJL, Bartels DM, Newell BR, Paolacci G, et al. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgm Decis Mak*. 2015;10(5):479–91.
24. Stein JS, Tegge AN, Turner JK, Bickel WK. Episodic future thinking reduces delay discounting and cigarette demand: an investigation of the good-subject effect. *J Behav Med* [Internet]. 2017 12 21; Available from: 10.1007/s10865-017-9908-1
25. Goodman JK, Cryder CE, Cheema A. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *J Behav Decis Mak*. 2013 7 2;26(3):213–24.
26. Huff C, Tingley D. “Who are these people?” Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*. 2015 7 1;2(3):2053168015604648.
27. Borodovsky JT, Marsch LA, Budney AJ. Studying Cannabis Use Behaviors With Facebook and Web Surveys: Methods and Insights. *JMIR Public Health Surveill*. 2018 5 2;4(2):e48. [PubMed: 29720366]
28. Cunningham JA, Godinho A, Kushnir V. Can Amazon’s Mechanical Turk be used to recruit participants for internet intervention trials? A pilot study involving a randomized controlled trial of a brief online intervention for hazardous alcohol use. *Internet Interv*. 2017 12;10:12–6. [PubMed: 30135748]
29. Tompkins DA, Huhn AS, Johnson PS, Smith MT, Strain EC, Edwards RR, et al. To take or not to take: the association between perceived addiction risk, expected analgesic response and likelihood of trying novel pain relievers in self-identified chronic pain patients. *Addiction*. 2018 1;113(1):67–79. [PubMed: 28645137]
30. Bickel WK, Stein JS, Moody LN, Snider SE, Mellis AM, Quisenberry AJ. Toward Narrative Theory: Interventions for Reinforcer Pathology in Health Behavior. In: Stevens JR, editor. *Impulsivity*. Springer International Publishing; 2017 p. 227–67. (Nebraska Symposium on Motivation).
31. Snider SE, LaConte SM, Bickel WK. Episodic Future Thinking: Expansion of the Temporal Window in Individuals with Alcohol Dependence. *Alcohol Clin Exp Res*. 2016 7;40(7):1558–66. [PubMed: 27246691]
32. Stein JS, Wilson AG, Koffarnus MN, Daniel TO, Epstein LH, Bickel WK. Unstuck in time: episodic future thinking reduces delay discounting and cigarette smoking. *Psychopharmacology* . 2016 10;233(21–22):3771–8. [PubMed: 27553824]
33. Strickland JC, Stoops WW. Feasibility, acceptability, and validity of crowdsourcing for collecting longitudinal alcohol use data. *J Exp Anal Behav*. 2018 7;110(1):136–53. [PubMed: 29873806]
34. Eisenberg IW, Bissett PG, Canning JR, Dallery J, Enkavi AZ, Whitfield-Gabrieli S, et al. Applying novel technologies and methods to inform the ontology of self-regulation. *Behav Res Ther*. 2018 2;101:46–57. [PubMed: 29066077]

35. Enkavi AZ, Eisenberg IW, Bissett PG, Mazza GL, MacKinnon DP, Marsch LA, et al. Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proc Natl Acad Sci U S A*. 2019 3 19;116(12):5472–7. [PubMed: 30842284]
36. Gleibs IH. Are all “research fields” equal? Rethinking practice for the use of data from crowdsourcing market places. *Behav Res Methods*. 2017 8;49(4):1333–42. [PubMed: 27515317]
37. Hara K, Adams A, Milland K, Savage S, Callison-Burch C, Bigham JP. A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM; 2018 p. 449:1–449:14. (CHI ‘18).
38. Irani LC, Silberman MS. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM; 2013 p. 611–20. (CHI ‘13).
39. Peer E, Vosgerau J, Acquisti A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav Res Methods*. 2014 12;46(4):1023–31. [PubMed: 24356996]
40. Berinsky AJ, Margolis MF, Sances MW. Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *Am J Pol Sci*. 2014 7 6;58(3):739–53.
41. Johnson MW, Bickel WK. An algorithm for identifying nonsystematic delay-discounting data. *Exp Clin Psychopharmacol*. 2008 6;16(3):264–74. [PubMed: 18540786]
42. Stein JS, Koffarnus MN, Snider SE, Quisenberry AJ, Bickel WK. Identification and management of nonsystematic purchase task data: Toward best practice. *Exp Clin Psychopharmacol*. 2015 10;23(5):377–86. [PubMed: 26147181]
43. Chandler J, Mueller P, Paolacci G. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav Res Methods*. 2014 3;46(1):112–30. [PubMed: 23835650]
44. Robinson J, Rosenzweig C, Moss AJ, Litman L. Tapped Out or Barely Tapped? Recommendations for How to Harness the Vast and Largely Unused Potential of the Mechanical Turk Participant Pool [Internet]. 2019. Available from:
45. Dupuis M, Meier E, Cuneo F. Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behav Res Methods*. 2019 10;51(5):2228–37. [PubMed: 30091086]
46. Waggoner P, Kennedy R, Clifford S. Detecting Fraud in Online Surveys by Tracing, Scoring, and Visualizing IP Addresses. *JOSS*. 2019 5 23;4(37):1285.
47. Leeper TJ. MTurkR: Access to Amazon Mechanical Turk Requester API via R. R package version 0. 2015;6(5.1).
48. Burleigh T. pyMTurkR [Internet]. 2019 9 Available from: <https://cran.r-project.org/web/packages/pyMTurkR/pyMTurkR.pdf>
49. McInnis B. Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers’ Experiences in Amazon Mechanical Turk. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016 <https://dl.acm.org/doi/10.1145/2858036.2858539>

Figure 1. The Growing MTurk Literature.

Figure 1 depicts the number of Google Scholar hits (y) over time (x) for the term “Amazon Mechanical Turk”, as of December 31, 2019.

Figure 2. Intervention validation through MTurk.

The top panel depicts delay discounting rates (A) and the number of cigarette puffs consumed (B) after engagement with either episodic future thinking (EFT) or a control condition (ERT), collected from an in-person study. The bottom panel depicts a summary measure of delay discounting rates (A) and cigarette purchase task data (B) after engagement with either episodic future thinking (EFT) or a control condition (ERT), collected via MTurk. Reprinted with permission.

Table 1.

Approaches to mitigate common concerns with Mturk

Concern	Approach
Non-naivete	<ul style="list-style-type: none"> • Post HITs in large batches, to increase the effective worker pool available. • Assign qualifications to Workers who have completed previous, similar HITs. Share these Worker IDs among collaborating research groups (43), with participant consent. • Allow Workers with a lower number of completed HITs to complete assessments (44). Note, however, that “reputations” are not granted until a Worker has completed >100 HITs.
Worker inattention	<ul style="list-style-type: none"> • Include attention check questions or “catch” trials. These can include instruction comprehension checks in addition to more traditional attention checks. • Full-screen tasks to minimize multi-tab browsing, as through commercial services (e.g., Inquisit Web) or custom-built web stimulus presentation tools. • Validated criteria for excluding responses or flagging data for inattention/inability to understand tasks (e.g., Johnson & Bickel criteria for excluding delay discounting task data). Ideally, these should be pre-registered, even informally, to reduce researcher degrees of freedom. Report full sample collected and reasons for data exclusion. • Screen based on Worker “reputation” (39). Note, however, that this common practice may reduce worker naivete (44).
Fraudulent responses	<ul style="list-style-type: none"> • Check for common VPS (virtual private server) signatures, such as the use of commonly “spoofed” IP addresses, post-hoc in collected data. • Check for response consistency and random responding (45) • Prevent multiple responses from matching IP addresses, or any responses from commonly fraudulent addresses (46)
Worker treatment	<ul style="list-style-type: none"> • Compensate appropriately for the actual duration of tasks. Report both planned and actual hourly compensation. • Compensate for lengthy screenings, considering that Workers estimate payment by the minute. • Exclude past participants through custom Qualifications, not by Blocking workers. • Provide an option for participants to withdraw consent while still being compensated, e.g. through custom survey codes.