Moeck Ella (Orcid ID: 0000-0002-5300-7316)

2

Food for thought: Commentary on Burnette et al. (2021) "Concerns and recommendations for using Amazon MTurk for eating disorder research"

Ella K. Moeck[1], Victoria M. E. Bridgland[2], & Melanie K. T. Takarangi[2]

[1]Melbourne School of Psychological Sciences, The University of Melbourne

[2]College of Education, Psychology, and Social Work, Flinders University

Corresponding author:
Ella Moeck (ellamoeck@gmail.com)
Melbourne School of Psychological Sciences
Redmond Barry Building
The University of Melbourne
Parkville 3010
Victoria, Australia

Abstract

Burnette et al. (2021) aimed to validate two eating disorder symptom measures among transgender adults recruited from Mechanical Turk (MTurk). After identifying several data quality issues, Burnette et al. abandoned this aim and instead documented the issues they faced (e.g., demographic misrepresentation, repeat submissions, inconsistent responses across similar questions, failed attention checks). Consequently, Burnette et al. raised concerns about the use of MTurk for psychological research, particularly in an eating disorder context. However, we believe these claims are overstated because they arise from a single study not designed to test MTurk data quality. Further, despite claiming to go “above and beyond” current recommendations, Burnette et al. missed key screening procedures. In particular, they missed procedures known to prevent participants who use commercial data centers (i.e., server farms) to hide their true IP address and complete multiple surveys for financial gain. In this commentary, we outline key screening procedures that allow researchers to obtain quality MTurk data. We also highlight the importance of balancing efforts to increase data quality with efforts to maintain sample diversity. With appropriate screening procedures, which should be pre-registered, MTurk remains a viable participant source that requires further validation in an eating disorder context.

Keywords: *online research, Mechanical Turk, data quality, open science, online recruitment platform, survey research*

Food for thought: Comment on Burnette et al. (2021) "Concerns and recommendations for using Amazon MTurk for eating disorder research"

Online recruitment platforms—including Amazon's Mechanical Turk (MTurk), CloudResearch, and Prolific—are increasingly popular because they offer fast, remote, and low-cost access to participants for psychological research. MTurk is the most popular of these platforms (Kennedy et al., 2020) and, as Burnette et al. note, is widely used by eating disorder researchers. Thus, Burnette et al. selected MTurk to validate two eating disorder symptom measures among transgender adults. After identifying several data quality issues (e.g., demographic misrepresentation, repeat submissions, inconsistent responses across similar questions, failed attention checks), the authors abandoned their original aim and instead documented the issues they faced. Based on their experience, Burnette et al. question the quality of data MTurk participants (termed "workers") provide, express doubts about prior MTurk eating disorder research, and suggest approaching future MTurk research with caution.

To our knowledge, Burnette et al. are the first to document MTurk data quality within eating disorder research and we commend their intentions. However, we believe Burnette et al.'s claims are overstated for several reasons. These claims are based on a single study not designed to test data quality. An evaluation of MTurk data quality should be designed with this goal in mind and systematically vary screening procedures (e.g., compare data when options to prevent low quality responses are, vs. are not, implemented; Eyal et al., 2021). It is also unclear whether the authors' claims that MTurk provides poor quality data are specific to the sample they were recruiting, or whether these claims generalize to other minority groups, psychopathological symptoms, and online data collection platforms (e.g., CloudResearch, Prolific). But the most notable issue we identified is that despite claiming to go "above and beyond" current recommendations for obtaining high quality MTurk data, Burnette et al. missed several critical recommendations. In short, while we agree with the call for increased transparency around data screening procedures, we argue that with appropriate procedures MTurk remains a valuable recruitment tool.

In 2018, psychological scientists were alarmed by a rapid drop in MTurk data quality, indicated by nonsensical answers to open-ended questions and duplicate GPS locations (Kennedy et al., 2020). Researchers initially assumed bots (i.e., software that runs automated scripts to complete tasks) were the cause, because of their predominance in the online world. But research has since shown that bots are relatively uncommon in samples sourced from online recruitment platforms; for example, Kennedy et al. (2020) found bots comprised 0.01% of their MTurk sample. We can instead attribute low quality data to participants who hide their true IP address by completing surveys through a Virtual Private Server (VPS) hosted in a commercial data center (i.e., "server farm"). People can operate multiple VPSs from a single computer, allowing them to complete surveys multiple times for financial gain. Unlike bots, *people* operate VPSs and can bypass checks designed to catch bots, like reCaptcha (i.e., a type of Turing test) and Honeypots (e.g., a question only a bot can see). Therefore, although these bot-checks might seem reliable at face value due to their familiarity in online contexts, they do not prevent "server farmer" responses. Moreover, because VPSs provide users with multiple unique IP addresses from a single device, attempts to block repeat respondents using IP addresses will not work. The use of VPSs likely explains Burnette et al.'s high rates of multiple completions and highlights the need to use additional data screening procedures to stop server farmers.

Below we outline several of these screening procedures, noting that no one procedure is a panacea and it is best to use a multilayered approach. Researchers should determine the details of any screening procedures *before* commencing data collection and report this information in the study pre-registration and any resulting publications.

To minimize the risk of server farmers, researchers should take steps to prevent multiple responses from the same person. Qualtrics—often used to host surveys or experiments—has a “prevent multiple submissions” option to stop respondents entering the survey more than once. But because users can circumnavigate this method by clearing cookies or using a private browsing mode, this option should be supplemented by IP-address based methods. Kennedy et al. (2020) recently developed code that checks respondents’ IP addresses against the IPHub database (https://iphub.info/) and flags potential VPS responses. This code can be embedded at the start of a Qualtrics survey to deny access to suspected server farmers. Online recruitment platforms offer similar services. If researchers wish to use MTurk, we recommend CloudResearch’s MTurk toolkit, which has options to block repeat MTurk IDs and suspicious geolocations (i.e., known server farm locations), and/or recruit from a large pool of participants (currently at 75,000 with new participants added weekly) who have passed attention and engagement measures. These types of options lead to similar quality data from MTurk (through CloudResearch) compared to Prolific (Eyal et al., 2021), which is a UK based online recruitment platform. MTurk itself offers a similar option for premium users.

In addition to IP address related procedures, respondents should demonstrate they come from the intended geographical population before starting the survey. For example, when seeking an English-speaking sample, researchers could include an English Proficiency Test toward the start of the survey. By setting up automatic scoring in Qualtrics, participants who do not reach the passing criteria (e.g., getting 7/10 correct) can be exited from the study. According to CloudResearch (https://www.cloudresearch.com/resources/blog/), these tests screen out around

70% of server farmers who typically come from unintended geographical populations (e.g., India; Kennedy et al., 2020). Of course, researchers should consider the ethical implications of this test, such as informing participants that passing is an eligibility requirement. Another option is to embed ‘cultural checks’ within the survey, like showing a picture of an eggplant and asking what it is called (US: “eggplant”; India: “brinjal”). CloudResearch states cultural checks screen out around 93% of server farmer responses. Finally, respondents should have to provide sensical responses to questions that require *detailed* open text answers (e.g., summarize task instructions in participants’ own words; Eyal et al., 2021). In response to complex open-ended questions, server farmers typically provide nonsensical one-word answers (e.g., “GOOD”, “nothing”), or copy answers verbatim from the question or from websites (e.g., Wikipedia). Although Burnette et al. screened for nonsensical open-ended responses (e.g., providing a numerical value for “occupation”), the questions Burnette et al. used likely required an insufficiently detailed answer to detect server farmers.

Given the aforementioned data quality procedures CloudResearch provides, we were surprised by Burnette et al.’s decision *not* to run their study through Cloud Research’s MTurk toolkit. While researchers must pay an additional fee to use this toolkit, increased data quality and fewer exclusions makes the fee worthwhile. For example, we estimate Burnette et al. would have saved approximately $3000 by using CloudResearch’s screening tools rather than paying for data they could not use. We were also surprised that Burnette et al. did not use multiple recruitment platforms, given their aim to recruit a sample of 2250 transgender adults who make up 0.4% of the US population. The authors rightfully identified the problem of potential misrepresentation when collecting “rare” data points (also see Agley et al., 2021), yet chose to use only MTurk and conduct the gender screening themselves. Prolific likely would have been an ideal platform for the intended research, given Prolific participants provide quality data (Eyal et al., 2021) and can be targeted on pre-provided demographic criteria. Before completing any studies, Prolific participants are asked "How do you describe yourself?" (male, female, trans male/trans man, trans female/trans woman, genderqueer/gender non-conforming, different identity, rather not say). Researchers can indicate during study setup whether they would like to target participants of a specific gender, reducing the likelihood that participants would lie about their gender to partake in a study.

Attention check items are another important consideration. Failing to screen for inattention may lead to spurious relationships between psychopathology measures (e.g., Sulik et al., 2021). But being *too stringent* on attention checks can be problematic, particularly when seeking participants who already fall into a small percentage of the population. Despite including two attention checks in the survey, Burnette et al. removed data from participants who failed a single attention check. This decision—which is similar to other MTurk research published in the *International Journal of Eating Disorders*—may be overly conservative for several reasons. First, people mind wander intermittently, meaning that inattention at one point does not necessarily mean inattention for the entire survey. Second, inattentiveness could be associated with psychopathology, such that overly strict criteria for inclusion based on attention checks may result in a sample that *under*-represents characteristics of interest (Agley et al., 2021). Conversely, this process may *over*-represent characteristics such as conscientiousness by creating a self-selection bias. To maintain sample diversity and data quality simultaneously, we suggest researchers include multiple checks of varying difficulty to capture general inattention. This recommendation fits with Prolific's guidelines, which do not allow researchers to exclude participants for failing a single attention check, except for surveys < 5-min. Third, we note that attention checks themselves may be imperfect and difficult for participants to understand. Thus, researchers should create *unique* but *fair* attention checks (for guidance, see https://researcher-help.prolific.co/hc/en-gb). Finally, we encourage researchers to consider MTurk participant naivety in addition to approval rating. MTurk workers who have completed many tasks (e.g., with 95+ HIT approval ratings) are most familiar with identifying attention checks and passing them, but do not necessarily provide quality data (e.g., Eyal et al., 2021).

Our research shows that these screening procedures yield quality data on disordered eating measures. In one study (McLean et al., 2021), we recruited vegan and omnivore participants from the US, UK, Canada, and Australia. Participants were sourced from MTurk (via CloudResearch's toolkit) and completed the same eating disorder symptom measures as Burnette et al.'s sample: the Eating Attitudes Test (EAT-26) and the Eating Disorder Examination Questionnaire (EDE-Q). However, before entering the survey participants had to score 7/10 on an English Proficiency Test. Participants who passed this test then answered a question about what food groups they eat/exclude. Participants who did not meet screening criteria for being vegan/omnivore were exited from the survey but paid a small amount for their time to reduce the likelihood of misrepresentation. Embedded within the main survey were three attention checks and a question requiring an open-ended response. We excluded participants who failed all three attention checks or provided nonsensical responses to an open-ended question. Participants showed excellent reliability on both the EAT-26 (alpha = .85) and the EDE-Q (alpha = .95), indicating good data quality (Eyal et al., 2021). For BMI, participants first chose their unit preference (height: feet-inches or cm, weight: pounds or kg), then provided their height and weight. Average BMI was in line with the US adult population, particularly for the omnivore participants (*M* = 26.2, vegan: *M* = 24.4), showing that giving participants the option to choose height/weight units helps to obtain interpretable and accurate height/weight data—as Burnette et al. suggest. Together, McLean et al. shows that data screening procedures designed to detect server farmers are effective in the context of (a) eating disorder research and (b) recruiting people who comprise a small percentage of the population (in this case, vegans).

Overall, we agree that screening procedures must be used when recruiting participants from online platforms. Eating disorder researchers should use these procedures as well as routinely check and report on the quality of any data sourced from platforms like MTurk. Indeed, obtaining high quality data is the first step toward validating MTurk for eating disorder research. In this commentary, we recommend several procedures (summarized here: https://osf.io/ezpfh/) that extend on Burnette et al.'s suggestions. We encourage researchers to look to studies outside the eating disorder field that comprehensively address MTurk data quality issues (e.g., Agley et al., 2021; Eyal et al., 2021; Kennedy et al., 2020). Finally, we echo Burnette et al.'s point that there should be greater transparency around screening procedures but suggest researchers go further: data screening procedures, and exclusions based on these procedures, should be disclosed in a study's pre-registration.

References

Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2021). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01665-8

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01694-3

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629. https://doi.org/10.1017/psrm.2020.6

McLean, C., Moeck, E. K., Sharp, G., & Thomas, N. A. (2021). Characteristics and clinical implications of the relationship between veganism and pathological eating behaviours. *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity*, 1–6. https://doi.org/10.1007/s40519-021-01330-1

Sulik, J., Ross, R. M., Balzan, R., & McKay, R. (2021). *Delusional Ideation and Data Quality: Are Classic Cognitive Biases an Artefact of Inattention?* (pp. 1–24) [Preprint]. https://psyarxiv.com/ntsve/