



Letters to the Editor

Re: Characteristics of a Mild Traumatic Brain Injury Sample Recruited Using Amazon's Mechanical Turk

To the Editor:

As Mr Bernstein and Dr Calamia note in their recent article, "Characteristics of a Mild Traumatic Brain Injury Sample Recruited Using Amazon's Mechanical Turk," Amazon's Mechanical Turk (MTurk) is an increasingly popular platform for recruiting research participants in the behavioral and health sciences [1]. We agree that MTurk provides efficient and low-cost access to high-quality data. However, researchers must also navigate important drawbacks to effectively use this platform. We use Bernstein and Calamia's study to highlight some of these drawbacks and offer strategies to address them.

A primary concern with all studies conducted online is the validity of participant responses. It is true that users of MTurk (workers) are no more likely to deceive researchers than participants collected through traditional methods [2]. However, when researchers recruit groups with low base rates in the population, such as people diagnosed with mild traumatic brain injury (mTBI), even a small frequency of invalid responses will have an outsized effect on results. The rate of fraud in a sample is determined by the proportion of deceptively eligible responses relative to the total number of eligible responses. For example, Chandler and Paolacci [3] demonstrated that when a small proportion (~3.5%) of an overall sample misstated that they had a child with autism, nearly half (45%) of the autism subsample was comprised of deceptive workers, due to the low base rate of autism in the population.

Bernstein and Calamia take the important step of concealing the purpose of the study in the task description (a step that some institutional review boards may not allow) and asking about multiple medical conditions in their screener. However, their workers could infer from the medical screener that one or more of the conditions listed was of interest

to the researchers, incentivizing them to answer affirmatively to as many as possible. Indeed, researchers have observed that an unusually high proportion of online participants select all possible alternatives when faced with potential screening questions [4]. Bernstein and Calamia do not provide data on the frequency of other conditions reported on their screener. The quality of their data would be stronger if workers who reported having all possible conditions, which would ideally include other rare or nonexistent conditions, were identified and excluded from participating.

A second important step to ensure quality is to minimize and monitor repeated attempts to complete a survey by the same person. Workers who are excluded from participating in a survey will sometimes reattempt it until they gain entrance. These workers will use knowledge gained in earlier attempts to qualify for the survey on later attempts [3]. Most survey software has technical methods that discourage repeated participation, but these are not turned on by default [3]. In addition, a substantial minority of respondents can defeat these features, so studies will ideally record identifying information like Amazon WorkerID's or IP addresses to link responses from the same person. It is also possible to limit fraud by concealing possible incentives. For example, a researcher could advertise and pay for the screener portion of the study and then either re-contact eligible workers or immediately route them to a second, "bonus" study that pays an additional sum [3,5].

Fraudulent responses can attenuate true relationships between variables. More worrisome, participants may try to infer expected associations between variables, leading fraudulent responses to produce false-positive results. For example, Wessling et al [5] found that people who (falsely) claimed to be elderly also claimed to eat more fiber. In this case, participants' theory was plausible but incorrect: people who were truly elderly made no such claim. Therefore, a fourth important step is to capitalize on this tendency by including measures of symptoms that the general public

D.S. Disclosure: nothing to disclose

J.C. Disclosure: nothing to disclose

assumes are prevalent but are in fact rare [6] and to pay close attention to discriminant validity more generally.

In sum, we agree that identifying and characterizing a sample of individuals with mTBI on MTurk is an important objective because access to this group has the potential to greatly accelerate the pace of discovery in nonclinical research. Bernstein and Calamia do many things right, but it is easy for readers unfamiliar with conducting online studies to overlook details that may impact data quality. Furthermore, the rarity of mTBI leaves a narrow margin for error. It is important for researchers using MTurk to understand its limitations and to ensure that they are taking all necessary steps to address them.

Danielle Shapiro, PhD

*Department of Physical Medicine and Rehabilitation
Michigan Medicine
Ann Arbor, MI*

Jesse Chandler, PhD

*Mathematica Policy Research
Institute for Social Research*

*The University of Michigan
Ann Arbor, MI*

<https://doi.org/10.1016/j.pmrj.2017.11.013>

References

1. Bernstein J, Calamia M. Characteristics of a mild traumatic brain injury sample recruited using Amazon's Mechanical Turk. *PM R* 2018;10:45-55.
2. Chandler J, Shapiro D. Conducting clinical research using crowd-sourced convenience samples. *Annu Rev Clin Psychol* 2016;12:53-81.
3. Chandler J, Paolacci G. Lie for a dime: When most prescreening responses are honest but most study participants are imposters. *Soc Psychol Personal Sc* 2017. <https://doi.org/10.1177/1948550617698203>. Available at journals.sagepub.com/doi/10.1177/1948550617698203. Accessed July 7, 2017.
4. Downes-Le Guin T, Mechling J, Baker R. Great results from ambiguous sources: Cleaning Internet panel data. ESOMAR World Research Conference: Panel Research 2006. Available at http://www.marketstrategies.com/user_area/content_media/Great_Results_From_Ambiguous_Sources.pdf. Accessed January 10, 2018.
5. Wessling KS, Huber J, Netzer O. Character misrepresentation by Amazon Turk Workers: Assessment and solutions. *J Consumer Res* 2017;44:211-230.
6. Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to study clinical populations. *Clin Psychol Sci* 2013;1:213-220.

Reply



We thank Drs Shapiro and Chandler for their helpful comments [1] on our manuscript [2] and their larger discussion of the potential pitfalls when using MTurk to assess clinical populations. In the space below, as a response to some of the concerns noted, we provide additional details not included in our manuscript on a few aspects of our study.

Shapiro and Chandler note the importance of being able to detect, and then exclude from participation, MTurk workers who falsely claim to have low-base-rate conditions in order to gain access to an assignment and ultimately to receive compensation. However, it should be noted that, relative to clinical conditions such as autism, which Shapiro and Chandler cite in their letter, mild traumatic brain injury (mTBI) has a rather high base rate. Previous epidemiological studies of traumatic brain injury (TBI) more broadly indicate a lifetime prevalence rate of between 5.7% and 8.5% [3,4], and more recent work indicates that more than one-fifth of adults have experienced a traumatic brain injury (TBI) with loss of consciousness. mTBI constitutes 90% of all TBIs, and many additional mTBIs go unreported [5]. As a

result, these figures likely reflect an underestimate of the lifetime prevalence of mTBI.

Given that the health screener was available to all MTurk workers meeting initial inclusionary criteria to complete, we were able to calculate a worker-reported mTBI prevalence rate in the MTurk population. Within our study, approximately 10% of workers who completed the screener met criteria for a previous mTBI, which seems plausible and not suggestive of a large number of individuals falsely claiming mTBI on our screening measure.

Shapiro and Chandler note that the quality of our data would be stronger if we excluded workers who endorsed all conditions on our health screener. We actually did do this as part of our study, as we excluded from participation all workers who endorsed a history of moderate-to-severe TBI on the screener. It is also worth noting that among participants included in our sample, less than 5% endorsed one-fourth (6) or more of the 23 conditions listed on the screener, less than 1% reported 10 or more conditions, and none endorsed more than 11. The vast majority of individuals endorsing more than one-fourth of conditions reported a number of emotional disorders (e.g., depression, anxiety-related disorders) that demonstrate high rates of comorbidity with one another [6], and prior research has shown that the prevalence of anxiety and depressive symptoms is as