International Journal of Social Research Methodology

ISSN: 1364-5579 (Print) 1464-5300 (Online) Journal homepage: http://www.tandfonline.com/loi/tsrm20

How serious is the 'carelessness' problem on Mechanical Turk?

Mara S. Aruguete, Ho Huynh, Blaine L. Browne, Bethany Jurs, Emilia Flint & Lynn E. McCutcheon

To cite this article: Mara S. Aruguete, Ho Huynh, Blaine L. Browne, Bethany Jurs, Emilia Flint & Lynn E. McCutcheon (2019): How serious is the 'carelessness' problem on Mechanical Turk?, International Journal of Social Research Methodology, DOI: 10.1080/13645579.2018.1563966

To link to this article: https://doi.org/10.1080/13645579.2018.1563966

Published online: 11 Jan 2019.

INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
https://doi.org/10.1080/13645579.2018.1563966

How serious is the 'carelessness' problem on Mechanical Turk?

Mara S. Aruguete[a], Ho Huynh[b], Blaine L. Browne[c], Bethany Jurs[d], Emilia Flint[e] and Lynn E. McCutcheon[f]

[a]Department of Social and Behavioral Sciences, Lincoln University, Jefferson City, MO, USA; [b]Department of Science and Mathematics, Texas A&M University, San Antonio, TX, USA; [c]Department of Psychology, Counseling, and Family Therapy, Valdosta State University, Valdosta, GA, USA; [d]Neuroscience Program, Transylvania University, Lexington, KY, USA; [e]Department of Behavioral Sciences, Black Hills State University, Spearfish, SD, USA; [f]North American Journal of Psychology, Winter Garden, FL, USA

ABSTRACT

This study compared the quality of survey data collected from Mechanical Turk (MTurk) workers and college students. Three groups of participants completed the same survey. *MTurk* respondents completed the survey as paid workers using the Mechanical Turk crowdsourcing platform. *Student Online* respondents also completed the survey online after having been recruited in class. Finally, *Student Paper-and-Pencil* respondents completed the survey on paper in a classroom setting. Validity checks embedded in the survey were designed to gauge participants' haste and carelessness in survey completion. MTurk respondents were significantly more likely to fail validity checks by contradicting their own answers or simply completing the survey too quickly. Student groups showed fewer careless mistakes and longer completion times. The MTurk sample tended to be older, more educated, and more ethnically diverse than student samples. Results suggest that researchers should pay special attention to the use of validity checks when recruiting MTurk samples.

ARTICLE HISTORY
Received 26 July 2018
Accepted 21 December 2018

KEYWORDS
MTurk; online sampling; college students; survey research; response set

In 2005, Amazon created Mechanical Turk (hereafter MTurk), a tool that permits temporary workers to complete computerized tasks for which they receive payment. Soon MTurk became a popular research tool because it allowed researchers to efficiently recruit a large and diverse sample at a low cost (Buhrmester, Talaifar, & Gosling, 2018; Desoto, 2016). Data for most survey studies can be collected in hours after posting, and automated data entry virtually eliminates data-entry errors (Berinsky, Huber, & Lenz, 2012). Short surveys can cost less than $.50 per participant, making large samples instantly accessible (Goodman, Cryder, & Cheema, 2013). Moreover, MTurk samples tend to be demographically heterogeneous (Berinsky et al., 2012). Since 2005, myths about online participants as social misfits were largely dispelled, and companies such as Qualtrics offered methodological tools to facilitate the collection of research data on the Internet (Buhrmester et al., 2018). The pool of workers has continued to swell and researchers have become increasingly engaged in MTurk participant recruitment (Goodman et al., 2013). Our study evaluates the use of MTurk for participant recruitment in social science research.

Early research reports about the use of MTurk were mostly positive. One study pointed out that MTurk has no risk of experimenter effects, and that the sample can be restricted to fit the needs of the researcher by the use of prescreening questions (Paolacci, Chandler, & Ipeirotis,

CONTACT Mara S. Aruguete aruguetem@lincolnu.edu Department of Social and Behavioral Sciences, Lincoln University, 820 Chestnut St., 310 FH, Jefferson City, MO 65102-0029

2010). Another found that inattention on the part of participants recruited through MTurk was about equal to that of college students in a face-to-face setting (Casler, Bickel, & Hackett, 2013). One widely read paper concluded that MTurk is fast, inexpensive, and as reliable as data obtained from traditional methods (Buhrmester, Kwang, & Gosling, 2011). Furthermore, MTurk participants are more demographically diverse (i.e., older, more educated, and more male) than typical college student samples (Berinsky et al., 2012; Buhrmester et al., 2011; Goodman et al., 2013; Paolacci et al., 2010). Another study also found that MTurk data are reliable, fast, and relatively inexpensive (Bates & Lanza, 2013). However, they recommended manipulation checks to determine if participants are paying attention because nine percent of their sample failed two simple checks ('I am having a heart attack right now' and 'I have no hands, eyes, or teeth'; Bates & Lanza, 2013). Yet another early study also recommended validity-check questions in order to quantify inattention levels and provide a reason for excluding data from those who are inattentive (Oppenheimer, Meyvus, & Davidenko, 2009).

More recent reports have emphasized a bit more caution. Chandler, Mueller, and Paolacci (2013) noted that some MTurk workers reported that they often respond to surveys while simultaneously watching television, listening to music, and interacting with others. Rouse (2015) found that the reliability estimate of an MTurk sample with no validity check was significantly lower than reliability estimates for the same scale reported by the scale developer (Goldberg, 1999). However, some respondents in Rouse's (2015) MTurk sample were asked if they had been honest and attentive in their responses. Deleting those who answered 'no' to these questions reduced the reliability gap. Desoto (2016) pointed out that the actual size of the MTurk population pool is smaller than previously believed, in part due to some workers who reportedly fill out hundreds of surveys every month. Further, he claimed, as did Krantz (2015) and Chandler et al. (2013) that there is no way to guarantee that workers are giving their undivided attention to survey items.

As a response to reports of inattentiveness, researchers have increasingly begun to use of embedded validity checks to assess the quality of data collected on MTurk. In an effort identify inattentive MTurk participants, Lian, Huynh, McCutcheon, Aruguete, and Murtagh (in press) included four items designed to detect contradictory responses. Their study excluded 13 percent of the original sample as a result of failing validity checks (Lian et al., in press), a proportion comparable with data screening procedures for some face-to-face data collection (Meade & Craig, 2012). Other studies have included additional MTurk validity checks such as identifying extremely unlikely response patterns or very short response times to complete the survey (Bernstein, Gillen, Aruguete, & McCutcheon, 2018). Since inattentiveness is often intermittent (Meade & Craig, 2012), using a variety of validity checks tends to increase the number of excluded participants, as different respondents often fail different checks. A very recent study (hereafter known as the Curiosity study; Aruguete et al., 2018) used multiple validity checks and found that 32% of respondents were excluded for inattentiveness (Aruguete et al., 2018). A high number of post-hoc deletions can become problematic because it may serve to reduce the diversity of the sample. Moreover, in an experimental design, excluding participants may introduce a confounding variable if the probability of exclusion is linked to experimental condition.

While inattentiveness is clearly an issue among MTurk workers, it remains unclear how MTurk samples compare with other participant samples. Differences in attentiveness might be related to the differing motivations of college students and MTurk workers. For example, college students might participate for a variety of reasons (e.g., course credit, loyalty to an instructor, or interest in the topic) while MTurk workers are primarily motivated by monetary compensation. One study (Goodman et al., 2013) compared MTurk workers with college students who either took the same survey online or face-to-face. The two college student groups scored similarly on a single validity/inattentiveness item, so their scores were combined. The students showed an attention failure rate of 11.5%, but the MTurk group had an attention failure rate of 33.8%. Some of this discrepancy was attributed to the larger number of MTurk workers for whom English was a second language.

However, the authors noted that, had they included data from those who failed the validity/inattentiveness item, their subsequent analyses of several personality measures would have produced only slightly different results. These differences would have mostly resulted in reduced levels of statistical significance (Goodman et al., 2013). On the other hand, some researchers have found no differences between MTurk workers and college student samples on attentiveness (Behrend, Sharek, Meade, & Wiebe, 2011; Casler et al., 2013; Hauser & Schwarz, 2015).

Shapiro, Chandler, and Mueller (2013) have encouraged researchers to continue to critically evaluate the appropriateness of recruiting MTurk samples. Some of the studies (Hauser, Sunderrajan, Natarajan, & Schwarz, 2016; Paolacci et al., 2010; Rouse, 2015) of carelessness among MTurk workers used only one either/or check for inattentiveness; thus, a truly inattentive MTurk worker had a 50% chance to pass. Meade and Craig (2012) recommend that data screening approaches simultaneously use a variety of attentiveness indicators, since careless responding tends to be sporadic, occurring on some questions or scales but not others. The vast majority of studies also fail to compare MTurk workers with the more commonly recruited sample of college students. Any sample will likely include some inattentive participants, but it is important to consider whether the MTurk platform or participant pool might increase the possibility of inattentiveness compared to other means of recruitment.

The present study

It is clear that some research participants will respond to surveys so swiftly that they arguably are paying little or no attention to item content. What is not clear is whether this undesirable behavior is more common among MTurk workers compared to college students faced with responding in one of two conditions. The present study compares survey data collected on MTurk respondents, a student online sample, and a student paper-and-pencil sample. We know of only one similar study (Goodman et al., 2013), which found that MTurk participants are less attentive to survey questions than are student samples. We therefore hypothesize that MTurk respondents will be more likely than student respondents to fail a variety of validity tests.

Method

Participants

A total of 340 participants completed the same survey. The quasi-experimental design compared three subsamples: MTurk participants, Student Online participants, and Student Paper-and-Pencil participants (See Table 1). The two student groups were recruited in classes taught by the authors. Students in these classes were randomly assigned to complete the survey using paper and pencil or online.

'MTurk participants' were 110 persons (18–34 years old) recruited through Amazon's MTurk. While the population of potential participants was international, all were registered with MTurk as English speakers. Most (72%) were college graduates, having earned a two-year degree (36), four-year degree (65), master's degree (4) or PhD or professional degree (1).

A second group of 129 participants designated 'Student Online' (18–51 years old), was recruited from four American universities located in the states of Georgia, Missouri, Kentucky,

Table 1. Demographic information about each subsample.

	n	Age (*SD*)	Male	Female	White	Asian	Black	Multi	Other
MTurk	110	24.45 (1.97)	63	47	45	43	13	2	7
Student Online*	129	21.71 (4.38)	32	95	69	1	45	10	4
Student P&P	101	21.46 (3.81)	26	73	57	2	30	4	7

N = 340; *two participants did not indicate gender

and South Dakota. While all were currently enrolled in college, 26% reported having already obtained a two-year (23), four-year (7), or master's degree (1).

A third group of 101 participants, designated 'Student Paper-and-Pencil' (18–44 years old), was recruited from the same American universities as the Student Online sample. Some participants (9%) had earned a two-year (4) or four-year (4) degree. The number of participants from groups two and three was approximately equal across all four universities. See Table 1 for more demographic information about each group.

An a priori power analysis using the GPower computer program (Erdfelder, Faul, & Buchner, 1996) indicated that a total sample size of 159 (assuming equal sample sizes) would be needed to detect a moderate effect size of Cohen's f = .25 (Cohen, 1988) with 80% power using a one-way ANOVA with an alpha at .05. Our sample size per group, while not equal across groups, still falls well within this parameter to ensure sufficient power of our experimental design.

Measures

Our measures consisted of the rate of respondents deleted from the Curiosity study because of carelessness in filling out a survey on which that study was based (Aruguete et al., 2018). The entire survey consisted of 75 items (including demographics) from the following scales; the Celebrity Attitude Scale (CAS), the Curiosity and Exploration Inventory (CEI-II), the Meaning in Life Questionnaire (MLQ), the Gratitude Questionnaire-Six (GQ-6), and the Psychological Flexibility Questionnaire (PFQ).

To minimize the likelihood of persons who were not seriously responding to our survey from contaminating our data analysis, we inserted validity checks. Two non-scored items (numbered 15 & 16), added to the end of the *Psychological Flexibility Questionnaire* (PFQ; Ben-Itzhak, Bluvstein, & Maor, 2014), were nearly mirror opposites of existing items numbered 5 and 10. For example, Item 15 read 'I do not feel ready to accept future changes' and item 10 read 'I feel ready to accept future changes.' Item 16 read 'I am not an open person in comparison with others,' but item 5 read 'I am an open person in comparison with others.' The 6-pt. response format ranged from 'not at all' to 'very much.' If respondents answered similarly (i.e., scored 1–2 or 5–6) on both 10 and 15 or they answered similarly on both items 5 and 16, we assumed they were not responding thoughtfully to the PFQ and we excluded them from the data analysis. In Table 2, we labeled this the PFQ contradiction.

We did something similar on the *Curiosity and Exploration Inventory-II* (CEI-II; Kashdan et al., 2009). Two non-scored items (numbered 11 & 12) added to the end of the CEI-II were almost exactly the opposite of existing items numbered 6 and 7. For example, item 11, 'I do not like to do things that are a little frightening,' was the opposite of item 6, 'I like to do things that are a little frightening.' Item 12, 'I rarely look for experiences that challenge how I think about myself and the world,' was the opposite of item 7, 'I am always looking for experiences that challenge how I think about myself and the world.' The response options were: Very slightly or not at all (1), A little (2), Moderately (3), Quite a bit (4), Extremely (5). If respondents answered similarly (i.e., scored 1–2 or 4–5) on items 6 and 11 or on 7 and 12, we assumed they were not responding thoughtfully to the CEI-II and we excluded them from the data analysis. In Table 2 we labeled this the CEI-II contradiction. We also deleted respondents that completed the entire survey in less than 150 seconds, our rationale being that it would be impossible to meaningfully and thoughtfully respond to such a large body of items and accompanying directions in less than two-and-a-half minutes.

Procedure

Permission to carry out this study was granted by the Institutional Review Boards at each participating university. The Curiosity Survey given to the MTurk group was presented through

Qualtrics, with order of the scales randomized to minimize the possibility of a systematic order effect. The order of questions within each scale did not differ. The survey was restricted to persons between the ages of 18 and 51. Qualtrics features a timer that allows researchers to determine exactly when workers finish a survey, and this time to completion was recorded for each respondent. MTurk workers were paid 75 cents each for participating.

Students were recruited to either the Student Online group or Student Paper-and-Pencil group via researchers from their respective universities. Researchers advertised an extra credit opportunity to participate in the survey and randomly assigned participants to one of the two different survey medium conditions: Online and Paper-and-Pencil. Each participant was awarded extra credit equal to one percent of their final course grade in a class taught by the respective author.

Students completed the Curiosity Survey (both online or paper-and-pencil) in university classrooms under the supervision of the author at each of the four respective universities. The Student Online group was treated almost exactly the same way as the MTurk group. They saw the same presentation through Qualtrics as the MTurk group and filled out the Curiosity Survey on a computer. The Student Paper-and-Pencil participants were timed by stopwatch. To reduce inaccuracy, students in the paper-and-pencil group participated in one of two or three small groups so that the researcher could more accurately record the finish time for each participant.

Completion times for the two computerized groups are assumed to be accurate. One person was eliminated from the Student Online group prior to analysis due to an extraneous three-hour recorded completion time, which greatly increased the duration variability in this group. Researchers at some institutions were required to include consent forms, whereas others were not. We made a 20-second adjustment for those participants who read consent forms. We also allotted five seconds in the paper-and-pencil groups between the time that participants finished and the time they handed the survey to the experimenter. We recognize that there are individual differences in the time taken for these adjustments (i.e., some participants glossed over the consent form while others actually read it carefully).

Results

From the MTurk sample of 110, we deleted a total of 51 respondents because they failed one or more validity checks. From the Student Online sample and the Student Paper-and-Pencil sample, we deleted 30 and 27 respondents respectively (see Table 2). We computed a chi-square for goodness of fit comparing the number of deleted and retained participants in each group. We obtained a significant result, X^2 (2, N = 340) = 16.31, $p < .001$, indicating that participants in the MTurk sample were more likely to be deleted for careless mistakes than the student samples. We did an additional one-way, between-subjects ANOVA comparing the three groups on the duration of time to complete the survey, $F(2, 337) = 15.53$, $p < .001$, partial $\eta^2 = .09$. Post hoc Tukey HSD tests indicated that the MTurk subsample (M = 383.09 sec.; SD = 189.78) completed the survey in about half the time of the Student Online subsample (M = 767.34; SD = 833.76; $p = .00$) and the Student Paper-and-Pencil subsample (M = 622.70; SD = 177.44; $p = .00$), showing that the

Table 2. Respondents deleted from the curiosity study.

Deletion Status	MTurk	Student Online	Student Paper-and-Pencil
Failed the CEI Contradiction	41	21	22
Failed the PFQ Contradiction	14	9	6
Failed Time-to-Completion Check	7	1	0
Total Deleted	51	30	27
Total Retained	59	99	74
Percent of Subsample Deleted	46%	23%	27%

Several respondents failed more than one validity check, especially in the MTurk group.

MTurk sample was particularly rushed in responding to the survey. However, there was no significant difference between the two student samples, $p = .10$.

All three subsamples were more likely to fail the CEI contradiction than the PFQ contradiction (See Table 2). This is likely related to the fact that the CEI is a 5-pt. scale (whereas the PFQ is a 6-pt. scale). If a participant was inattentive, they had a greater probability of answering similarly on contradictory items when there were fewer response options available.

Discussion

It seems quite clear from Table 2 that MTurk respondents were more careless than respondents from the two student groups, who were not statistically different from each other. The difference between the MTurk group and the two student groups on failed time-to-completion was not huge (7 to 1 to 0). However, considering that the MTurk workers, on average, took about half the time to complete as the two student groups, it could be argued that the criterion for excluding thoughtless respondents on the basis of time-to-completion was not very stringent. It could well be that much more than 150 seconds were needed to *thoughtfully* respond to all the items in the Curiosity study. If the criterion time were set at 180 seconds, the difference between the MTurk group and the two student groups diverges even further (14 to 3 to 0). At 210 seconds, the ratio becomes 20 to 3 to 0. Therefore the MTurk group appears to be far more hurried completion of the survey than the student samples. These data suggest that time-to-completion is an important validity check to consider when sampling MTurk respondents.

Our results contradict the findings of some researchers (Behrend et al., 2011; Casler et al., 2013; Hauser & Schwarz, 2015), but they are consistent with the results of Goodman et al. (2013), in that MTurk respondents showed more carelessness than both groups of college students. Our high failure rate as compared with most previous studies is probably attributable in part to the fact that we had three validity checks, whereas others had only one (Goodman et al., 2013; Hauser et al., 2016; Paolacci et al., 2010; Rouse, 2015) or two (Bates & Lanza, 2013). We join with others (Bates & Lanza, 2013; Lian et al., in press; Oppenheimer et al., 2009; Rouse, 2015) in recommending the use of validity checks in survey research, especially for data collected through MTurk.

Why were MTurk respondents more careless than college students in our study? One possibility was suggested by Desoto (2016), who claimed that some MTurk workers fill out hundreds of surveys a month. If this is so, they may become jaded and more careless in their haste. While there are no available data on the number of surveys a typical college student fills out, informal observation of student behavior by the authors leads us to speculate that they arguably fill out fewer than a dozen over a four-year span.

We believe that anonymity might be another reason for the differences we found. Although our survey did not require any overtly identifying information, MTurk respondents are likely to know that researchers will never learn their identities, so they can be careless without much fear of suffering negative consequences. MTurk does have an option to allow researchers to reject incomplete or otherwise unacceptable responses. When MTurk workers' acceptance rates fall below 99 percent, they fail to qualify for certain opportunities as a worker. However, it is generally not a sound practice for researchers to reject and withhold payment from participants because participants should be able to withdraw from a study at any time without penalty. Instead, researchers may elect to add the respondent to a special list in their MTurk account, which they can use to exclude that respondent for future studies. This measure does not help other researchers avoid this respondent, but at least a researcher can prevent the same irresponsible respondent from entering all of their studies. On the other hand, students are given assurances that none of their survey responses will result in any recrimination, but they are also aware that they are known by the professor who is conducting the research, and they may be cautious about doing anything that might have a negative effect on their course grade. Students may also be motivated to help a professor by carefully completing a survey. Nevertheless, about one in every

four students was not particularly careful in the Curiosity study. This is a concerning finding given the dominance of student samples in psychological research.

We are reminded of the famous obedience studies conducted by Milgram (1974), and the finding that more obedience to the demands of the experimenter occurred when participants were face-to-face with the experimenter, rather than at a remote distance. In a sense, filling out a survey *thoughtfully* and *carefully* is a way of complying with the implicit expectations of a researcher who is either in the same room or nearby. MTurk workers, like the remote distance participants in one of Milgram's studies, probably felt less compelled to comply by giving the Curiosity survey their full attention. That said, this distance-quality difference may be quite small when remote and face-to-face participants are drawn from the same population, as were our two student samples.

If ethnic diversity is a goal for researchers, especially if Asian participants are sought, then MTurk might be a good way to collect data, providing validity checks are used. We had 43 Asians in our MTurk sample, as opposed to 3 in our student samples combined. The greater number of Asian participants results from the international pool of MTurk workers. Favorable currency exchange rates might also be a motivating factor for some international participants.

Our study had a few limitations. The MTurk group was significantly older, more educated, and more ethnically diverse than our two student samples. Clearly our groups were not comparable on some variables of potential importance. On the other hand, we did not control for these differences because we wanted to compare the student samples with an MTurk sample that was fairly typical of MTurk samples in general. We believe we succeeded, inasmuch as previous MTurk samples have been found to be more educated (Paolacci et al., 2010), and more diverse (Buhrmester et al., 2011; Goodman et al., 2013; Paolacci et al., 2010) than non-MTurk samples. Despite random assignment of student subsamples, our student online group tended to be more educated than the student paper-and-pencil group. This may reflect varying compliance rates between the two subsamples.

A potential limitation with some recent MTurk studies is an increased presence of 'bots.' Bots are MTurk users who employ special technology to circumvent screening methods so that they can profit from submitting multiple responses to the same survey in order to collect additional payment. A simple counter measure to this problem is to use open-ended questions because bots provide extremely poor quality responses to such questions. Another solution would be to include reCAPTCHA in the survey (for more details and other suggestions see: Dennis, Goodson, & Pearson, 2018). Because we included open-ended questions, bots were not an issue for our study, but this problem is important to note for researchers intending to use MTurk in their research.

Future research on this topic might attempt to replicate the present study by restricting the age, educational level, ethnicity, and nationality of an MTurk sample, with the goal of making it more comparable to a typical college student sample. Nonetheless, our findings suggest that researchers should be aware of the limitations of recruiting samples via MTurk. While use of MTurk might be a convenient, quick, and low-cost way of attaining a diverse, international sample, researchers should carefully consider including multiple validity checks.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Mara S. Aruguete is a professor at Lincoln University of Missouri. Dr. Aruguete received her PhD in Comparative and Physiological Psychology at the University of California, Davis in 1994. Her research focuses on celebrity attitudes and the impacts of race/ethnicity in higher education.

Ho Huynh is an assistant professor of psychology at Texas A&M University-San Antonio. Dr. Huynh received his Ph.D. in Social/Health Psychology at the University of California, Riverside in 2014. Dr. Huynh conducts research on the topics of health, sports, and humility.

Blaine L. Browne is an associate professor at Valdosta State University. Dr. Browne received his PhD in Experimental Psychology at Oklahoma State University in 2002. His research focuses on cognition, learning, and memory.

Bethany Jurs is an assistant professor at Transylvania University. She received her Ph.D. in Cognitive Neuroscience and Cognitive Science at Indiana University in 2009. Her research concerns the neural correlates of learning, fingerprint examination, and facial recognition.

Dr. Emilia Flint is an associate professor at Black Hills State University. She received her Ph.D. in Counseling Psychology from the University of North Dakota in 2010. Dr. Flint has research interests in the field of sport, exercise, and performance psychology.

Lynn E. McCutcheon is the Founder and Editor in Chief of the North American Journal of Psychology. He is a retired faculty member of DeVry University in Orlando, Florida. His research interests include the psychological characteristics of celebrity worshippers and sports psychology.

References

Aruguete, M. S., Huynh, H., McCutcheon, L. E., Browne, B., Jurs, B., & Flint, E. (2018). Are curiosity, meaning in life, and psychological flexibility linked to admiration for celebrities? Manuscript submitted for publication.

Bates, J. A., & Lanza, B. A. (2013). Conducting psychology student research via the Mechanical Turk crowdsourcing service. *North American Journal of Psychology*, *15*(2), 385–394.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavioral Research*, *43*, 800–813.

Ben-Itzhak, S., Bluvstein, I., & Maor, M. (2014). The psychological flexibility questionnaire (PFQ): Development, reliability and validity. *WebmedCentral Psychology*, *5*(4), 1–10. WMC004606.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351–368.

Bernstein, M, Gillen, M, Aruguete, M. S, & McCutcheon, L. E. (2018). Disconnection from nature and the admiration of celebrities. Australian Journal of Environmental Education.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*(2), 149–154.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*, 2156–2160.

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaivete among Amazon MTurk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods*, *46*, 112–130.

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

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). MTurk workers' use of low-cost "Virtual Private Servers" to circumvent screening methods: A research note. *Social Science Research Network*. doi:10.2139/ssrn.3233954

Desoto, K. A. (2016, March). Under the hood of Mechanical Turk. *Observer*. Retrieved from https://www.psychologicalscience.org

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1–11.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of MTurk samples. *Journal of Behavioral Decision Making*, *26*, 213–224.

Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on attention checks than do subject pool participants. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-015-0578-z

Hauser, D. J., Sunderrajan, A., Natarajan, M., & Schwarz, N. (2016). Prior exposure to instructional manipulation checks does not attenuate survey context effects driven by satisficing or Gricean norms. *Methods, Data, Analyses*, *10*(2), 145–220.

Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D., & Steger, M. F. (2009). The curiosity and exploration inventory-II: Development, factor structure, and psychometrics. *Journal of Research in Personality*, *43*, 987–998.

Krantz, J. H. (2015). Can the World Wide Web be used for research? In M. A. Gernsbacher & J. R. Pomerantz (Eds.), *Psychology and the real world* (pp. 10–17). New York, NY: Worth.

Lian, B., Huynh, H., McCutcheon, L. E., Aruguete, M. S., & Murtagh, M. (in press). Is gambling addiction related to celebrity addiction? Journal of Projective Psychology and Mental Health

McCutcheon, L. E., Aruguete, M., McCarley, N. G., & Jenkins, W. J. (2016). Further validation of an indirect measure of celebrity stalking. *Journal of Studies in Social Sciences*, *14*(1), 75–91.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455.

Milgram, S. (1974). *Obedience to authority*. New York, NY: Harper & Row.

Oppenheimer, D. M., Meyvus, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304–307.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013, January 31). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*, 213–220, Online.