Graduating from Undergrads:
Are Mechanical Turk Workers More Attentive than Undergraduate Participants?

Colin A. Capaldi[1]*

[1] Department of Psychology, Carleton University, Ottawa, Ontario, Canada

* Corresponding author
E-mail: colin_capaldi@carleton.ca

Abstract

With the advent of the internet and crowdsourcing sites like Amazon's Mechanical Turk, psychologists and other social scientists are increasingly going online to recruit participants for their studies. Although websites like Mechanical Turk provide novel opportunities for speedy and inexpensive data collection from more diverse samples, many researchers are concerned that workers may be less attentive and provide lower quality data compared to participants who are recruited from other sources. Given these concerns and the mixed findings from previous research, the current investigation tested whether Mechanical Turk workers are more attentive and effortful while completing psychological studies than undergraduate students. Based on data from the recent large, collaborative Many Labs 3 project, it was found that Mechanical Turk workers report paying more attention and exerting more effort than undergraduate students. Mechanical Turk workers were also more likely to pass an instructional manipulation check than undergraduate students. Based on these results, it appears that concerns over participant inattentiveness may be more applicable to samples recruited from traditional university participant pools than from Mechanical Turk.

Introduction

For most of psychology's history, undergraduate students have been the prototypical participant used in research attempting to uncover truths about the human condition [1, 2, 3]. Despite long-standing and repeated concerns over the representativeness of undergraduate samples and the generalizability of findings based on data from them [4, 5, 6], undergraduate participant pools have remained a convenient and popular source for recruiting experimental subjects.

With the advent of the internet and crowdsourcing sites, however, psychologists are increasingly going online to recruit participants [7]. One online labor market that has received a considerable amount of attention and popularity among academics is Amazon's Mechanical Turk (MTurk). In addition to being used by businesses to crowdsource the completion of a wide range of small tasks over the internet, MTurk is also being utilized by social scientists to recruit participants for online studies. With more than half a million individuals from around the world signed up as workers on the site [8], the pool of potential participants is much larger and more diverse than traditional recruitment sources offer [9, 10]. Moreover, data collection with MTurk tends to be relatively quick and inexpensive; most individuals are willing to work for less than a few dollars an hour [11] and studies that may have taken months to collect data for in the past can be completed in mere hours or days [9]. Given all these advantages, it is not surprising that MTurk has rapidly gained popularity as a useful tool for conducting research on a variety of topics.

Nevertheless, there are concerns that the aforementioned benefits of MTurk may come at a cost. In an informal survey, participant inattentiveness and low data quality were listed as the greatest concerns of using MTurk for two-thirds of researchers [12]. Empirical investigations of whether these worries are justified have found mixed results. Some studies have found that MTurk workers are less likely to carefully read through instructions and pass instructional manipulation checks [13] than student samples [14]. Relatedly, engaging in potentially distracting activities while completing surveys appears to be more common among MTurk workers than one would hope [12]. In contrast, researchers have been able to successfully replicate established psychological effects using samples from MTurk (e.g., [10, 15]) and some research has found that MTurk workers are as likely or even more likely to pass instructional

manipulation checks than traditional undergraduate samples [10, 16].

Given researcher concerns and mixed findings in this area, the purpose of the current investigation was to examine whether MTurk workers differ from undergraduate students in attentiveness and effortfulness when participating in psychological studies. The data analyzed herein allowed for high-powered comparisons of MTurk workers to undergraduate students from twenty different institutions. Building off of the research in this area, both self-report and behavioral measures of attention and effort were analyzed to test whether perceptions about the potential drawbacks of running studies on MTurk are accurate.

Methods

Data and participants

Publicly available data from the Many Labs 3 project [17] were analyzed to investigate the research question of interest. This collaborative, crowdsourced project was primarily interested in whether the detectability of psychological effects and the characteristics of participants vary over the academic semester. Taking place over 30 minutes, participants completed a series of experimental tasks and individual difference questionnaires. The study was run in twenty different university labs (N = 2,696) and online with a sample from MTurk (N = 737) during the 2014 fall semester.

American and Canadian undergraduate students were recruited from each university's participant pool and were offered course credit for participating. They were required to come to the lab to participate, although most of the study took place on a computer. Of the undergraduate students who reported their gender, 30.02% were male. The age of participants ranged from 13 to 54 years (M = 19.30, SD = 2.67).

MTurk workers were recruited as a comparison sample and were compensated $1.25 for

participating in the online study. Besides having to be from the United States, no other eligibility restrictions (i.e., based on experience or reputation) were set. Of the MTurk workers who reported their gender, 51.40% were male. The age of participants ranged from 18 to 72 years (M = 35.11, SD = 10.89).

Materials

Self-reported attention was measured by asking participants to rate how closely they paid attention to the instructions and experimental tasks on a 5-point scale ranging from 1 (*none*) to 5 (*I gave the tasks my undivided attention*). Self-reported effort was assessed by asking participants to rate the amount of effort they put into the experimental tasks on a 5-point scale ranging from 1 (*no effort*) to 5 (*I tried my hardest*). Not surprisingly, these two self-report items were positively correlated, $r(3203) = .54$, $p < .001$. In an attempt to minimize socially desirable responding to these questions, participants were told that their ratings would not affect the compensation they would receive for participating.

An instructional manipulation check was included as a behavioral measure of attention. Participants were presented with a question that was seemingly interested in their leisure activity preferences. Following a long paragraph of instructions was an item which read “In my free time I prefer” and six response options that included “engaging in hobbies”, “watching TV, reading, music”, “being in nature”, “exercising”, “cooking or eating”, and “other”. The last option had a textbox that allowed participants to type in their own response. The last two sentences of the preceding paragraph, however, told participants that the researchers were interested in whether they were actually reading the directions, and asked them to ignore the other response options and write “I read in the instructions” in the textbox. If participants ignored the other response options and wrote “I read the instructions” (or something similar) in the textbox, they passed the

instructional attention check; if not, they failed.

Ethics Statement

Ethical approval from each university's institutional review board was obtained before data was collected. Participants were presented with an informed consent form and gave their written consent to participate in this study.

Results

An independent samples *t*-test was conducted to examine whether the MTurk and undergraduate samples differed in self-reported attention. The test was statistically significant, $t(3206) = 20.44$, $p < .001$, $d = .94$, with MTurk workers reporting that they paid more attention during the study ($M = 4.60$, $SD = 0.58$) than undergraduate students ($M = 3.93$, $SD = 0.74$). The independent samples *t*-test remained statistically significant when outliers (i.e., self-reported attention scores ± three standard deviations from the mean) were excluded from the analysis, $t(3204) = 20.47$, $p < .001$, $d = .94$. As the distribution of self-reported attention scores were not normally distributed, a Mann-Whitney *U* test was also performed. Similar to previous parametric analyses, it was also statistically significant, $U = 388178.00$, $p < .001$. As one can see in Fig 1, the MTurk sample had higher overall self-reported attention than all of the undergraduate samples.

To examine whether the MTurk and undergraduate samples differed in self-reported effort, an independent samples *t*-test was conducted. It was statistically significant, $t(3203) = 19.36$, $p < .001$, $d = .89$, with MTurk workers reporting that they exerted more effort ($M = 4.40$, $SD = 0.77$) than undergraduate students ($M = 3.71$, $SD = 0.78$). Results were similar when outliers were excluded from the analysis, $t(3197) = 19.59$, $p < .001$, $d = .89$, and when a Mann-Whitney *U* test was conducted due to non-normality, $U = 404138.00$, $p < .001$. Fig 2 shows that

overall self-reported effort was lower in all of the undergraduate samples.

If participants are being honest when reporting their levels of attention and effort, one would expect to see reliable differences in those who passed versus failed the instructional manipulation check. This is, in fact, what was observed when additional independent sample *t*-tests were conducted. Regardless of whether the participant was an undergraduate student or MTurk worker, self-reported attention was higher for those who passed the instructional manipulation check ($M = 4.19$, $SD = 0.71$) than those who failed it ($M = 3.78$, $SD = 0.78$), $t(3181) = 14.25$, $p < .001$, $d = .56$. Similarly, self-reported effort was higher among individuals who passed the instructional manipulation check ($M = 3.94$, $SD = 0.82$) compared to those who did not ($M = 3.62$, $SD = 0.77$), $t(3178) = 10.54$, $p < .001$, $d = .40$. Interpretation of the results remained the same for self-reported attention, $t(3179) = 14.17$, $p < .001$, $d = .55$, and self-reported effort, $t(3172) = 10.24$, $p < .001$, $d = .39$, when outliers were excluded. Results were also similar for self-reported attention, $U = 1411996.50$, $p < .001$, and self-reported effort, $U = 1331739.50$, $p < .001$, when nonparametric statistics were used. In sum, these results provide some evidence for the convergent validity of the self-report items; they appear to be an, at least somewhat, accurate reflection of the attentiveness and effortfulness of participants.

Finally, a chi-square test of homogeneity was conducted to examine whether the likelihood of passing or failing the instructional manipulation check differed for MTurk and undergraduate samples. The test was statistically significant, χ^2 (1, $N = 3200$) = 219.27, $p < .001$, with 93.96% of the MTurk workers passing the instructional manipulation check but only 62.23% of undergraduate passing it. See Fig 3 for the percentage of individuals who passed the instructional manipulation check at each data collection site.

Discussion

Despite seemingly reasonable concerns about using MTurk to recruit participants for research, results from this investigation suggest that MTurk workers, on average, pay more attention and exert more effort than undergraduate students while participating in psychological studies. This was not only found with self-report measures of attentiveness and effortfulness, but with a behavioral measure as well. Beyond being statistically significant, the differences between participants from the two recruitment sources tended to be large in magnitude [18]. These large differences were observed even when no eligibility restrictions beyond location were used when recruiting workers from MTurk, potentially offering a more equal comparison than if the MTurk sample was restricted to solely individuals with high reputations and lots of previous experience.

The difference in the percentage of individuals who passed the instructional manipulation check was quite considerable, with almost all of the MTurk workers but less than two-thirds of the undergraduate students passing it. This superior and extremely high pass rate for MTurk workers is consistent with some of the most recent research comparing them to undergraduates [17], which includes data from the first Many Labs project [19]. As argued elsewhere [17], MTurk workers may be passing attention checks at higher rates because they come from a non-replenishing participant pool that is frequently exposed to these types of checks and incentivized to pay attention and expend more effort as their worker reputation and compensation can sometimes depend on their performance [12].

Regardless of why MTurk workers are especially adept at passing instructional manipulation checks, the results offer continued concern for the less than ideal passing rates and the relative lack of attention to detail among individuals recruited from traditional university participant pools. Even with requiring them to complete the study in a lab setting with an experimenter nearby, more than one-third of undergraduate students failed the instructional

manipulation check. This might be especially concerning for social psychologists and other researchers whose subtle experimental manipulations might be missed due to participant inattentiveness (see [13]). In addition to including data quality indicators in studies with samples from MTurk, it appears like it would be wise to include them when doing research with undergraduate students as well.

To my knowledge, this investigation is the first to examine differences between MTurk workers and undergraduates in self-reported attention and effort. Although higher scores could arguably be attributed to MTurk workers' greater desire to please researchers [20] and present themselves in a favorable light [21], these self-report items did differ reliably between those who passed and those who failed the instructional manipulation check. This suggests that participants were answering these questions in an at least somewhat honest and accurate manner, and that differences between recruitment sources are not completely attributable to confounding variables like social desirability. Regardless of whether attention and effort were measured behaviorally or by self-report, a clear advantage for MTurk workers over undergraduates emerged.

There are several limitations, notes of caution, and areas for future research that should be mentioned. Although the Many Labs 3 data allowed for high powered comparisons of MTurk workers and undergraduates, the samples were geographically restricted to the US and Canada. The pattern of results found in the current investigation may not necessarily generalize to samples recruited from outside of these countries. For instance, the passing rate would likely be lower if no location restrictions were used when recruiting MTurk workers as performance on instructional manipulation checks partially depends on one's language proficiency (i.e., non-native speakers are more likely to fail [14]). Similarly, the findings might also not generalize to individuals recruited from other crowdsourcing sites (see [22]). In the current experimental

design, the instructional manipulation check was included near the end of the study. Differences in rates of passing the instructional manipulation check might be less pronounced, albeit still large, if it was presented earlier as one study found that MTurk workers are slightly more likely to pass these types of attention checks when they are presented near the end versus the beginning of the study [17]. The length of the study session and the amount of compensation given to workers may also be important moderators. Finally, although the pool of potential participants on MTurk is quite large, a large percentage of workers report being familiar with common experimental paradigms such as the prisoner's dilemma [12] and this non-naïveté has the potential to influence subsequent participant responses and study results (e.g., [23]). Thus, researchers should carefully consider this issue when deciding on which measures, experimental paradigms, and eligibility requirements to use; at the very least, researchers should attempt to assess whether workers have previously participated in similar studies [12].

Conclusion

In sum, the current investigation adds to the growing body of research showing that MTurk workers may actually be more attentive and effortful when completing surveys than researchers initially thought. In fact, results suggest that many concerns about inattentiveness may be more applicable to individuals from traditional undergraduate participant pools than workers from MTurk. Along with the greater efficiency and relative inexpensiveness of collecting data, this study provides one more benefit for graduating from undergrads and recruiting from MTurk.

Acknowledgements

Thank you to all those involved in the Many Labs 3 project. This paper would not have

been possible without all of their hard work and commitment to open, replicable science.

References

1. Arnett JJ. The neglected 95%: why American psychology needs to become less American. Am Psychol. 2008 Oct;63(7):602–14. doi: 10.1037/0003-066X.63.7.602

2. Peterson RA. On the use of college students in social science research: insights from a second-order meta-analysis. J Consum Res. 2001 Dec;28(3):450–61. doi: 10.1086/323732

3. Wintre MG, North C, Sugar LA. Psychologists' response to criticisms about research based on undergraduate participants: a developmental perspective. Can Psychol. 2001 Aug;42(3):216–25. doi: 10.1037/h0086893

4. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci. 2010 Jun;33(2-3):61–83. doi: 10.1017/S0140525X0999152X

5. Myers DG. Social psychology. 5th ed. New York: McGraw-Hill; 1983.

6. Smart RG. Subject selection bias in psychological research. Can Psychologist. 1966 Apr;7(2):115–21. doi: 10.1037/h0083096

7. Sargis EG, Skitka LJ, McKeever W. The internet as psychological laboratory revisited: best practices, challenges, and solutions. In: Amichai-Hamburger Y, editor. The social net: the social psychology of the internet. UK: Oxford University Press; 2014. pp. 253–70.

8. Amazon Mechanical Turk [Internet]. Seattle: Amazon.com, Inc; c2005-2015 [cited 2015 Jul 15]. Available: https://requester.mturk.com/tour

9. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci. 2011 Jan;6(1):3–5. doi: 10.1177/1745691610393980

10. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judg Decis Mak. 2010 Jun; 5(5):411–19. doi: 10/10630a/jdm10630a

11. Horton JJ, Chilton, LB. The labor economics of paid crowdsourcing. Proceedings of the 11th ACM Conference on Electronic Commerce; 2010 June 7-11. Available: http://arxiv.org/pdf/1001.0627.pdf

12. Chandler J, Mueller P, Paolacci G. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. Behav Res. 2013 Jul;46(1):112–30. doi: 10.3758/s13428-013-0365-7

13. Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: detecting satisficing to increase statistical power. J Exp Soc Psychol. 2009 Jul;45(4):867–72. doi: 10.1016/j.jesp.2009.03.009

14. Goodman JK, Cryder CE, Cheema A. Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. J Behav Dec Making. 2013 Jul;26(3):213–24. doi: 10.1002/bdm.1753

15. Crump MJC, McDonnell JV, Gureckis TM. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLOS ONE. 2013 Mar;8(3):e57410. doi: 10.1371/journal.pone.0057410

16. Hauser DJ, Scwarz, N. Attentive Turkers: MTurk participants perform better on online attention checks than subject pool participants. Behav Res Methods. 2015 Mar; In press. doi: 10.3758/s13428-015-0578-z

17. Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, et al. Many Labs 3: evaluating participant pool quality across the academic semester via replication; 2015. Pre-print. Available: osf.io/ct89g

18. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum. 1988.

19. Klein RA, Ratliff KA, Vianello M, Adams Jr. RB, Bahník Š, Bernstein MJ, et al. Investigating variation in replicability: a "many labs" replication project. Soc Psychol. 2014;45(3):142–52. doi: 10.1027/1864-9335/a000178

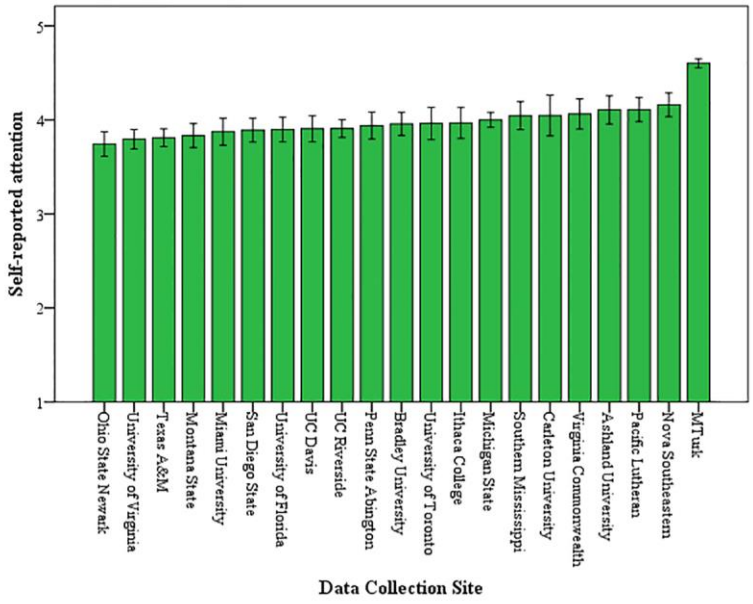
20. Paolacci G, Chandler J. Inside the Turk: understanding Mechanical Turk as a participant pool. Curr Dir Psychol Sci. 2014 Jun;23(3):184–8. doi: 10.1177/0963721414531598

21. Behrend TS, Sharek DJ, Meade AW, Wiebe, EN. The viability of crowdsourcing for survey research. Behav Res Methods. 2011 Sep;43(3):800–13. doi: 10.3758/s13428-011-0081-0

22. Peer E, Samat S, Brandimarte L, Acquisti A. Beyond the Turk: An empirical comparison of alternative platforms for crowdsourcing online behavioral research; 2015. Pre-print. Available: http://papers.ssrn.com/abstract=2594183

23. Rand DG, Peysakhovich A, Kraft-Todd GT, Newman GE, Wurzbacher O, Nowak MA, et al. Social heuristics shape intuitive cooperation. Nat Commun. 2014 Apr; 5. doi: 10.1038/ncomms4677

328 **Fig 1. Mean self-reported attention across data collection sites.**
329 Error bars are 95% confidence intervals.
330



331

Fig 2. Mean self-reported effort across data collection sites.
Error bars are 95% confidence intervals.

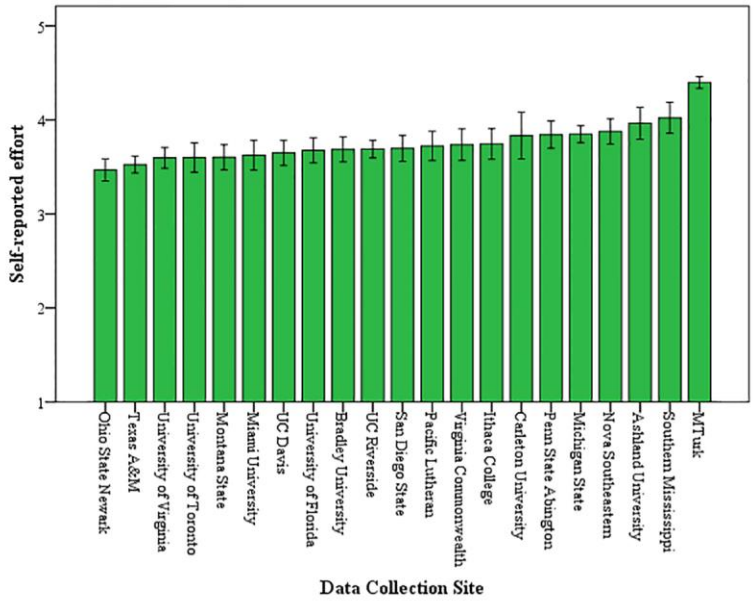


Fig 3. Percentage of individuals who passed/failed the instructional manipulation check at each data collection site.