2019

Reliability of MTurk Data from Masters and Workers for Personality and Cognitive Abilities

Steven V. Rouse
Pepperdine University, steve.rouse@pepperdine.edu

Reliability of MTurk Data from Masters and Workers for Personality and Cognitive Abilities

Steven V. Rouse

Pepperdine University

Author Note

Steven V. Rouse, Social Sciences Division, Pepperdine University.

This research was supported by an internal grant from Pepperdine University's Seaver Research Council. This research was approved by the Seaver College Institutional Review Board (protocol #17-11-654) and followed the ethical standards of the American Psychological Association. There were no conflicts of interest in conducting this research. Studies 1 and 2 were presented at the 2019 conference of the Society for Personality and Social Psychology.

Correspondence concerning this article should be addressed to Steve Rouse, Social Sciences Division, Pepperdine University, Malibu CA 90263. E-mail: steve.rouse@pepperdine.edu.

ABSTRACT

Previous research has supported the use of Amazon's Mechanical Turk (MTurk) for online data collection in individual differences research. Although MTurk Masters have reached an elite status because of strong approval ratings on previous tasks (and therefore gain higher payment for their work) no research has empirically examined whether researchers actually obtain higher quality data when they require that their MTurk Workers have Master status. In two different online survey studies (one using a personality test and one using a cognitive abilities test), the psychometric reliability of MTurk data was compared between a sample that required a Master qualification type and a sample that placed no status-level qualification requirement. In both studies, the Master samples failed to outperform the standard samples.

Keywords

Amazon's Mechanical Turk; MTurk; Online Data Collection; Psychometrics; Score Reliability

Reliability of MTurk Data from Masters and Workers for Personality and Cognitive Abilities

Introduction

Amazon's Mechanical Turk (MTurk; Amazon, 2011) has quickly become a valuable resource for psychological researchers (Buhrmester, Talaifar, & Gosling, 2018). As an Internet-based platform that allows "Requestors" to hire "Workers" for brief Human Intelligence Tasks (HITs), researchers have often used MTurk as a means of gaining survey data for online research on personality and individual difference variables. A PsycINFO search conducted in March 2019 using "MTurk" or "Mechanical Turk" as subject terms resulted in the identification of one publication in 2010, six in 2011, 17 in 2012, 34 in 2013, 59 in 2014, 125 in 2015, 192 in 2016, 258 in 2017, and 333 in 2018, demonstrating the speed with which empirical psychological research has incorporated this data acquisition resource. Goodman and Paolacci (2017) highlighted several strengths of MTurk data for research purposes that may be reasons for this rapid growth. First, MTurk data is less expensive than data gained through many other methods, allowing for more democratization in the research process (as researchers who do not have large budgets are able to be involved in high-quality research) and for more exploratory and more confirmatory research than can be conducted using more expensive means. Second, MTurk samples (representing a wide range of ages, socioeconomic backgrounds, racial and ethnic groups, and countries of origin) tend to be more diverse than samples obtained using many traditional methods, allowing researchers to focus specifically on subgroups that might be hard to study using traditional data collection methods. Third, MTurk provides a high level of flexibility, allowing for cross-cultural research, longitudinal research, and even research in which Workers engage in online real-time interactions with others. Fourth, we can presume high levels of quality for MTurk data because of its built-in incentive structure; since Workers who submit poor

quality work can be rejected or given low approval ratings (which can then prevent a Worker from being eligible for additional HITs) Workers are incentivized to be attentive to detail and conscientious in their work.

Balanced against the benefits of using MTurk for data collection purposes, Buhrmester et al. (2018) noted several factors that could have a negative effect on the quality of the data. For example, plausible concerns might be raised about inattentiveness while participating in research, dishonesty in affirming qualifications to participate in studies for which one is not actually qualified, and the non-naiveté and familiarity with research measures that results when a person completes a large number of research studies.

Although Goodman and Paolacci (2017) presented the incentive structure as a reason for confidence in the quality of MTurk data, one might reasonably wonder if the opposite were true. It is plausible that because Workers are aware that a Requester might reject their work or provide low approval ratings, there may be an implicit pressure to provide socially desirable responses, and this impression management could affect the validity of a measure. In addition, if people are paid for task completion, this might increase the likelihood of rushing through tasks, resulting in inattentiveness. As a general rule, this does not appear to be the case. For example, Ramsey, Thompson, McKenzie, and Rosenbaum (2016) showed a greater level of attentiveness to research instructions for an MTurk sample than for an undergraduate sample. Likewise, Thomas and Clifford (2017) found that inattentiveness in an experimental procedure was no more problematic for MTurk samples than for other convenience samples, and MTurk samples were found to be engaged in experimental procedures and to be committed to producing high-quality data. Thus, as a general rule, the external incentivization of MTurk payment does not appear to lead to poor or invalid data.

Research on the quality of MTurk data has especially focused on the question of score reliability, or the consistency of the data. Coefficients exceeding .70 have been documented for MTurk data for temporal consistency (Holden, Dennie, & Hicks, 2013; Kim & Hodgins, 2017), inter-rater consistency (Tosti-Kharas & Conley, 2016), and internal consistency (Bates & Lanza, 2013; Buhrmester, Kwang, & Gosling, 2011; Miller, Crowe, Weiss, Lynam, & Maples-Keller, 2017). For most of these MTurk reliability studies, the documentation of reliability took the form of contrasting reliability estimates obtained for MTurk samples against those obtained for standard samples. For example, Johnson and Borden (2012) calculated a Cronbach's alpha of .86 for a measure of empathy taken in a face-to-face lab setting and a Cronbach's alpha of .83 when this same test was taken by an MTurk sample. Rouse (2015) used a hypothesis testing procedure developed by Bonett (2003) to evaluate the statistical significance of the differences in reliability estimates for MTurk data and traditional data. Rouse showed that the inclusion of "best practices" (such as asking respondents to Opt-in or Opt-out of the study) resulted in reliability estimates for MTurk samples that were not statistically different from those calculated for community samples.

Although a large number of articles have documented the psychometric quality of MTurk responses, very little research has been conducted to evaluate the effect of a "best practice" advocated by Amazon (2011)—hiring MTurk Masters rather than non-Master Workers. "Master" refers to a designation given to an MTurk Worker who has demonstrated superior performance on a large number of HITs for several Requesters. Requesters can choose to limit their HITs to only being completed by MTurk Masters; presumably, this would result in high-quality data because "Masters must maintain this high level of performance or risk losing this distinction" (Amazon, 2011, p. 8). However, there is an additional financial cost; Amazon

charges a 5% fee for HITs that are designated exclusively for Masters (above and beyond the 40% fee charged for all HITs). “Because Masters have demonstrated accuracy, they can command a higher reward for their HITs” (p. 8). Lovett, Bajaba, Lovett, and Simmering (2017) offered a supportive rationale based on Equity Theory; since Masters receive higher payment than non-Master Workers, one may presume that they will put forth more effort and more attention into the research task. However, no studies have empirically examined whether or not Masters provide more accurate or reliable survey data.

A literature search only identified one publication that addressed the quality of data from MTurk Masters. Lovett et al. (2017) conducted a survey of 40 MTurk Masters, asking both open-ended and forced-choice questions about MTurk compensation, their motivation while completing MTurk HITs, the settings in which they complete MTurk HITs, and their perceptions of the quality of their responses. In general, these MTurk Masters believed that their work was high-quality, but that it could be affected by factors such as the fairness of compensation and attention-checks that encourage careful responding. While the responses given by these MTurk Masters are aligned with Amazon’s (2011) expectation that they provide especially accurate data, Lovett et al. only sought self-report perceptions of their own work and did not attempt to empirically contrast the quality of work produced by Masters and non-Master Workers.

The present studies sought to begin exploring this deficit in the research literature, to determine whether Masters generate more reliable data and, if so, whether the increased quality justifies a higher research expense. For the present studies, the quality of the data was assessed based on internal consistency reliability; this decision is aligned with Viswanathan’s (2005) taxonomy of measurement error and the methods best suited for identifying such error. For example, Viswanathan noted that internal consistency estimates are well-suited for identifying

generic random error that may be present within the administration of a measure. Therefore, if one subject sample is more attentive to a survey than is another sample (and therefore obtains scores that have less random measurement error), this difference in attentiveness and data quality will likely be seen in the internal consistency estimates of reliability. Thus, Amazon's claim of superior data produced by Masters could be evaluated by statistically comparing the reliability estimates obtained for the same measures when restricting data collection to Masters and when foregoing such a restriction. In the present studies (one using a measure of personality and one using a measure of cognitive ability), independent data sets were collected from MTurk Masters and non-Master Workers, with internal consistency reliability estimates calculated for each sample. The hypothesis was that Masters would generate more reliable data than non-Master Workers; support for this hypothesis would justify a higher payment for these elite survey-respondents.

Reliability for MTurk Masters and MTurk Workers on a Personality Test

The purpose of the first study was to determine whether data obtained from MTurk Masters were more reliable than data obtained from non-Master Workers on a personality measure. The data collection and data analysis plans were preregistered at [Masked for review]. The analysis strategy was to collect two independent data sets and calculate Cronbach's alphas for each. Using Bonett's (2003) process for statistically evaluating the difference between reliability estimates, a one-tailed hypothesis was set with a .05 *p*-value: The Null Hypothesis was that the Masters' data were not more reliable than the non-Masters' data, and the Alternate Hypothesis was that the Masters' data were more reliable. The study was approved by the [Masked for review] Institutional Review Board prior to data collection.

Method

Materials. Two nearly identical surveys were created, with informed consent information and the questions themselves administered online in a single window within the MTurk platform. To follow the principle of Open Materials advocated by the Open Science Collaboration (2015), the survey has been publicly archived at [Masked for review]. Pilot testing of the survey suggested that it would be reasonable to expect that this 30-item survey could be completed in 10 minutes. Based on this, a payment of $1.50 was selected (for an effective wage of $9.00/hour, which exceeded the $8.25/hour median US state minimum wage at the time of data collection).

Demographic Information. Both versions of the survey began with questions to assess age, gender, and race/ethnicity. The version of the survey that was not restricted to Masters also included a question to assess whether or not the respondent held Master status.

Agentic and Communal Values Scale (ACV; Trapnell & Paulhus, 2012). With 24 Likert-type items, the ACV provides independent measures of two main dimensions of personality and motivation—Agency (i.e., "getting things done" and "getting ahead") and Communion (i.e., "getting along"). Respondents are asked about the importance of 24 values as guiding principles in their lives, with 12 items such as "Achievement (reaching lofty goals)" and "Status (high rank, wide respect)" for Agency and 12 items such as "Altruism (helping others in need)" and "Compassion (caring for others, displaying kindness)" for Communion. Respondents were asked to indicate the salience of each value to their lives on a 9-point scale from 1 (i.e., "Not at all important to me") to 9 (i.e., "Extremely important to me"). Trapnell and Paulhus reported internal consistency reliability estimates of .83 for both scales. This measure was selected for this present study because it is relatively brief and yet it provides scale scores for two distinctly different personality traits. It was also a viable scale because of consistently high reliability estimates across a wide range of samples. Finally, it was selected because a literature

search conducted in January 2017 suggested that none of the publications citing Trapnell and Paulhus (2012) included the terms “MTurk” or “Mechanical Turk” in their abstracts. This leads to the assumption that MTurk samples would not have been exposed often to these questions (unlike some measures that are frequently used in MTurk studies), making it unlikely that the respondents would be influenced by familiarity with the measure.

Opt-in/Opt-out Question. The final question on the survey followed the recommendation of Rouse (2015), asking respondents to indicate whether or not their data should be included in the analyses. Subjects were assured that their response was confidential, it would not affect the MTurk approval rating they would receive, and it would not affect their payment. They were asked to either Opt-In (i.e., “You should keep my data; I paid attention and answered honestly”) or Opt-Out (i.e., “You should delete my data; honestly, I wasn’t really taking this seriously.”) from inclusion in the data analyses.

Master Sample. To be eligible for inclusion in the first sample, MTurk Workers had to be in the United States and have earned Master status. With these criteria set, responses were collected from 80 respondents; this sample size was selected to meet to Bonett’s (2003) recommendation for testing the statistical significance of the difference between two reliability estimates (specifying a one-tailed significance of .05, power of .80, and an anticipated range of reliability estimates from .70 to .85). All of the respondents answered the final Opt-In/Opt-Out question by indicating that their responses should be included for data analysis. Ages ranged from 22 years to 63 years (M = 36.88, SD = 9.18). The sample included 50 men and 30 women, and race/ethnicity self-identifications included European American (n = 67), Latinx/Hispanic (n = 9), African American (n = 6), and Asian American (n = 4); the total number of race/ethnicity

self-identifications exceeded 100% because respondents were allowed to select multiple self-identifications.

Worker Sample. To be eligible for the second sample, MTurk workers had to be in the United States and not be included in the first sample. All repondents indicated in the Opt-In/Opt-Out question that their data should be included in analyses. Of the 80 respondents, 51 were men and 29 were women, and the race/ethnicity self-identifications included European American (n = 56), African American (n = 11), Asian American (n = 8), Latinx/Hispanic (n = 7), Native American (n = 1), and Other (n = 1). Their ages ranged from 18 to 74 (M = 34.11, SD = 11.80). The majority (n = 71) did not have MTurk Master status[1].

Results and Discussion

To follow the principle of Open Data advocated by the Open Science Collaboration (2015), the survey data have been publicly archived at [Masked for review].

The two samples did not differ on mean scores for either of the two scales. On the overall score for Agency, the Masters (M = 54.81, SD = 16.88) and the non-Master Workers (M = 59.08, SD = 16.23) were not significantly different (t = 1.61, p = .11, d = 0.26). Similarly, on the overall score for Communion, the Masters (M = 76.23, SD = 17.60) and the non-Master Workers (M = 77.34, SD = 17.39) were not significantly different (t = 0.90, p = .69, d = 0.06). The correlations between Agency and Communion scales for the Masters (r = -.01, p = .92, 95% CI = [-.23, .21])

[1] A small proportion of the sample (n = 9) did have Master status; however, excluding these participants from the sample would result in data that do not reflect typical MTurk research. Although many MTurk studies do not restrict inclusion to MTurk Masters, a literature search conducted in January 2017 could not identify any MTurk studies that specifically excluded Masters, and the MTurk system itself only permits excluding non-Masters. By comparing responses from a Masters-only sample with a sample that did not specify status, this study was able to examine the effect of following a process that Amazon considers a "best practice" (i.e., Masters-only data collection).

and the non-Master Workers (r = -.07, p = .54, 95% CI = [-.29, .15]) suggests that these two scores are largely uncorrelated, which is consistent with the two-factor structure proposed by Trapnell and Paulhus (2012).

The reliability estimates did not differ significantly across samples for either scale. On the Agency scale, the Cronbach's alpha for the Masters was .89, and the Cronbach's alpha for the non-Master Workers was .87. The difference (Z = 0.68, one-tailed p = .25) did not meet criteria to reject the Null Hypothesis. On the Communion scale, the Cronbach's alpha for the Masters was .91, and the Cronbach's alpha for the non-Master Workers was .90. The difference (Z = 0.67, one-tailed p = .25) did not meet criteria to reject the Null Hypothesis. Thus, the results of the first study did not suggest that MTurk Masters produce data with higher levels of reliability on a well-established personality measure.

Reliability for MTurk Masters and MTurk Workers on a Cognitive Ability Test

The purpose of the second study was to determine whether data obtained from MTurk Masters on a measure of cognitive ability were more reliable than data obtained from MTurk non-Master Workers. The data collection and data analysis plans were preregistered at [Masked for review]. The analysis strategy was to collect two independent data sets and calculate a Cronbach's alpha for each. Using Bonett's (2003) process for statistically evaluating the difference between reliability estimates, a one-tailed hypothesis test was set with a .05 p-value: The Null Hypothesis was that the Masters' data were not more reliable than the non-Master Workers' data, and the Alternate Hypothesis was that the Masters' data were more reliable. The study was approved by the [Masked for review] Institutional Review Board prior to data collection.

Method

Materials. Two nearly identical surveys were created to be administered online with informed consent and survey questions on a single window within the MTurk platform. Because the usage agreement for the cognitive ability test used in this study prohibits researchers from presenting items in a publicly accessible location, this study does not follow the Open Materials principle advocated by the Open Science Collaboration (2015). Pilot testing of the survey suggested that it would be reasonable to expect that this 22-item survey could be completed in 13 minutes. Based on this, a payment of $2.00 was selected (for an effective wage of $9.23/hour, which exceeded the $8.25/hour median US state minimum wage at the time of data collection).

Demographic Information. Both versions of the survey began with questions to assess age, gender, and race/ethnicity. The version of the survey that was not restricted to Masters also included a question to assess whether or not the respondent held Master status.

International Cognitive Ability Resource (ICAR)—Verbal Reasoning (Condon & Revelle, 2014). The ICAR is a set of public domain cognitive ability tests, assessing a range of abilities similar to those included on many tests of intelligence. The Verbal Reasoning scale includes 16 multiple-choice items that assess knowledge of word meanings, logical relationships, and general factual knowledge. Condon and Revelle reported an internal consistency reliability estimate of .76 for data collected during scale development. This scale was selected for the present study because (despite its brevity) it is a very demanding measure of cognitive ability. Because all items are text-based (unlike some of the visual items included in other ICAR scales), it lends itself well to be used for an online survey. Finally, it was selected because a literature review did not yield any instances in which MTurk data were collected using this measure, suggesting that it was unlikely for MTurk Workers to have a high level of familiarity with the measure (which could lead to inflation of scores through the practice effect).

Opt-in/Opt-out Question. The surveys ended with the same Opt-In/Opt-Out question used for Study 1.

Master Sample. The 80 respondents in the Masters sample met the qualification criteria (i.e., being in the United States and having been grated Master status); all indicated in the Opt-In/Opt-Out question that their data should be included in analyses. The 44 men and 36 women had ages ranging from 23 to 62 (M = 36.78, SD = 8.70), and self-identified as European American (n = 63), Latinx/Hispanic (n = 9), African American (n = 7), and Asian American (n = 4).

Worker Sample. The 80 respondents in the Worker sample were in the United States and none had participated in the previous survey. All respondents indicated in the Opt-In/Opt-Out question that their data should be included in analyses. The 50 men and 30 women had ages ranging from 18 to 63 (M = 32.78, SD = 9.77), though one miscoded his age and two opted not to provide a response. Respondents self-identified as European American (n = 68), Asian American (n = 6), African American (n = 5), and Latinx/Hispanic (n = 1), though one participant opted not to self-identify. The majority (n = 73) did not have Master status.

Results and Discussion

To follow the principle of Open Data advocated by the Open Science Collaboration (2015), the survey data have been publicly archived at [Masked for review].

The mean scores on the ICAR Verbal Reasoning scale did not differ significantly (t = 1.75, p = .08, d = 0.28) between the Masters (M = 10.50, SD = 3.46) and non-Master Workers (M = 9.65, SD = 2.61).

The reliability estimate obtained for the Masters (Cronbach's alpha = .67) was much lower that the reliability estimate obtained for the non-Master Workers (Cronbach's alpha =

.81)[2]. Because a one-tailed significance test was performed to test whether Masters' data are more reliable, the difference (Z = -2.33, one-tailed p = .99) did not meet the criteria to reject the Null Hypothesis[3]. Thus, the results of the second study did not suggest that MTurk Masters produce data with higher levels of reliability on a well-established cognitive ability measure.

General Discussion

Previous research has been supportive of the use of MTurk for research data collection. Although there are some differences that can be expected between MTurk samples and general community samples, a wide range of psychological phenomena have been replicated using MTurk samples, and many studies have documented that the psychometric quality of the data is as strong as would be expected for more traditional means of data collection.

Amazon (2011) recommended hiring MTurk Masters (with an additional incremental fee) if one seeks higher quality data. However, no research had empirically examined this claim prior to the present project. In this project, two different studies failed to support this claim, with no evidence of higher reliability for MTurk Masters relative to data sets that did not require Master status. In fact, contrary to Amazon's claim, the reliability of data provided by Masters was

[2] Although Condon and Revelle (2014) reported a coefficient alpha of .76, it is not possible to contrast this with the reliability estimates obtained for either sample in this study because Condon and Revelle's estimate was based on an amalgamation of responses from 34,229 respondents; however, subjects in their sample completed various subset of the items, not the full scale. The significance test to contrast coefficient alphas suggested by Bonett (2003) requires specification of the sample size completing the full test, which is not possible to determine for the Condon and Revelle data. Therefore, statistical comparison of either sample to the value reported by Condon and Revelle is not possible.

[3] If a two-tailed significance test had been performed, the results would have been strong enough to reject the Null Hypothesis. However, a one-tailed test was used in order to evaluate the claim by Amazon (2011) that hiring MTurk Masters is a "best practice" that enhances the likelihood of higher quality data.

substantially lower than data provided by a general MTurk sample for one of the two studies. Additional research should be conducted to determine whether this finding is replicable.

If researchers gather additional replication data to explore the reliability differences between Masters' data and non-Master Workers' data, it is very possible that the results will continue to be contradictory. After all, in the present project one study yielded no clear difference in the psychometric quality between the two samples, while the other suggested that (if anything) the Masters' data were psychometrically weaker. If this pattern of contradictory results continues, the results should be evaluated within the context of research by Hamby and Taylor (2016) on the effect of "survey satisficing" and "survey optimizing" on the quality of MTurk data. Optimizing refers to instances of careful and effortful consideration of questions before providing "the best" responses, whereas Satisficing refers to situations in which a respondent provides quick, "good enough" responses. Hamby and Taylor demonstrated that these response patterns may affect the quality of MTurk data, such that Optimizing may be more common among MTurk respondents when motivation is high and task difficulty is low, but that Satisficing may be more common among MTurk respondents as task difficulty rises or motivation falls. In the present project, reliability levels were very similar on the personality measure (which is a relatively effortless survey to complete), but the reliability levels differed for the cognitive ability measure (which poses a much greater cognitive demand on the respondent). Perhaps a complex interaction effect exists between MTurk Master/Non-master status and Optimizing/Satisficing. If MTurk Workers seek a Master status, it is plausible that they may put an emphasis on producing high quality work regardless of the cognitive demand of the HIT; it is plausible, however, that MTurk Masters might simply be satisfied with producing "good enough" work when the cognitive demand of the HIT gets high.

In addition, as noted by Lovett et al. (2018), MTurk Masters acknowledge high levels of familiarity and expertise with self-report surveys and acknowledge working quickly on research measures in order to maximize compensation. These same Masters reported the belief that the quality of their data was strong, and acknowledged being observant of possible "attention checks" that might affect their compensation and approval ratings. Nevertheless, it is possible that Masters (many of whom are working as a primary source of income) become more inattentive than they realize when completing complex tasks like cognitive abilities tests. If this is the case, on some tasks Masters might actually provide lower quality data relative to non-Master Workers due to test-wiseness and the desire to finish work quickly. However, substantial additional research will be needed to determine whether or not this pattern replicates.

Although the present studies did not find higher reliability estimates for data obtained by Masters, undermining Amazon's (2011) claim that the employment of Masters is a "best practice", additional psychometric statistics could provide information about other aspects of data quality. For example, factor analytic procedures can assess the consistency of the structure of a measure and tests of validity can assess the existence of systematic cross-construct error (Viswanathan, 2005). Central to the question of the quality of data obtained by Masters, additional research could examine the number of missing responses, or the length and complexity of answers to free-response open-ended survey items; within the MTurk environment, both of these could be affected by a Worker's concern for having a HIT approved and receiving a positive evaluation. However, while the present studies did not refute Amazon's claim, the burden of responsibility lies with Amazon if the company wishes to promote hiring Masters as a "best practice".

References

Amazon. (2011). *Requester best practices guide.* Retrieved from http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf.

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

Bonett, D. G. (2003). Sample size requirements for comparing two alpha coefficients *Applied Psychological Measurement, 27,* 72–74. doi: 10.1177/0146621602239477.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3-5. doi: 10.1177/1745691610393980

Buhrmester, M., 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,* 149 - 154. doi: 10.1177/1745691617706516

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43,* 52 – 64. doi: 10.1016/j.intell.2014.01.004

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research, 44,* 196 – 210. doi: 10.1093/jcr/ucx047

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: A experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement, 76,* 912 – 932. doi: 10.1177/0013164415627349

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's Mechanical Turk. *Computers in Human Behavior, 29,* 1749 – 1754. doi: 10.1016/j.chb.2013.02.020

Johnson, D. R., & Borden, L. A. (2012). Participants at your fingertips: Using Amazon's Mechanical Turk to increase student-faculty collaborative research. *Teaching of Psychology, 39,* 245–251. doi: 10.1177/0098628312456615.

Kim, H. S., & Hodgins, D. C. (2017). Reliability and validity of data obtained from alcohol, cannabis, and gambling populations on Amazon's Mechanical Turk. *Psychology of Addictive Behaviors, 31,* 85 – 94. doi: 10.1037/adb0000219

Lovett, M., Bajaba, S, Lovett, M., & Simmering, M. J. (2017). Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology: An International Review*, *66,* 1 – 28. doi: 10.1111/apps.12124

Miller, J. D., Crowe, M., Weiss, B., Lynam, D. R., & Maples-Keller, J. L. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. *Personality Disorders: Theory, Research, and Treatment, 8,* 26 – 34. doi: 10.1037/per0000191

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* 943–951. doi: 10.1126/science.aac4716

Ramsey, S. R, Thompson, K. L., McKenzie, M., & Rosenbaum, A. (2016). Psychological research in the Internet age: The quality of web-based data. *Computers in Human Behavior, 58,* 354 – 360. doi: 10.1016/j.chb.2015.12.049

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior, 43,* 304 – 307. doi: 10.1016/j.chb.2014.11.004

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77,* 184 – 197. doi: 10.1016/j.chb.2017.08.038

Tosti-Kharas, J, & Conley, C. (2016). Coding psychological constructs in text using Mechanical Turk: A reliable, accurate, and efficient alternative. *Frontiers in Psychology, 7,* 1 – 9. doi: 10.3389/fpsyg.2016.00741

Trapnell, P. D., & Paulhus, D. L. (2012). Agentic and communal values: Their scope and measurement. *Journal of Personality Assessment, 94,* 39 – 52. doi: 10.1080/00223891.2011.627968

Viswanathan, M. (2005). *Measurement error and research design.* Sage: Thousand Oaks, CA, USA.