MISS Measurement Instruments for the Social Sciences

PsychOpen GOLD

Advances in Methodology

# Master Turkers: An Assessment of Data Quality

Christopher Trenge[1], James D. Griffith[1]

**[1]** *Department of Psychology, Shippensburg University, Shippensburg, PA, USA.*

Measurement Instruments for the Social Sciences, 2024, Vol. 6, Article e13619, https://doi.org/10.5964/miss.13619

**Received:** 2024-01-02 • **Accepted:** 2024-04-18 • **Published (VoR):** 2024-07-03

**Handling Editor:** Michael Bosnjak, Trier University, Trier, Germany

**Corresponding Author:** Christopher Trenge, Department of Psychology, The University of Alabama, Box 870348, Tuscaloosa, AL 35487-0348, USA. E-mail: ctrenge@gmail.com

**Supplementary Materials:** Data, Materials [see Index of Supplementary Materials]

## Abstract

Amazon's Mechanical Turk has greatly increased in popularity in recent years considering recent world events as well as due to the increased acceptance of technology in the field of research. Because of this, it is essential that the research methods associated with conducting research online be evaluated. The present study evaluated if Amazon's upper echelon of workers, Master Turkers, provide a higher quality of data relative to workers without that designation. This was evaluated using two scales that are validated and have been extensively used in research. The results showed that Master Turkers were found to have worse performance on scales (lower reliability) compared to non-Master Turkers. This data highlights an issue that potential researchers should be aware of when using the Mechanical Turk, as well as problem that should be addressed by Amazon.

## Keywords

Mechanical Turk, AMT, Master Turker, data quality

Access to pools of human research participants is often paramount to conducting research across many different fields across a wide variety of topics. During the past decade or so, the use of technology has been used to access groups that were inaccessible in the past. Due to recent global circumstances, as well as general acceptance of technology as a viable source for research, there has been an increase in research being conducted online. According to De Man et al. (2021), the necessitation of collecting data online due to the inability to conduct in person research as a result of the Covid-19 pandemic has resulted in a dramatic increase in the usage of online surveys. When considering what platform to use for online data collection, there are several options.

## Mechanical Turk

One popular option for data collection is Amazon's Mechanical Turk (AMT; Amazon, 2018). AMT is an online work platform that allows users (Requestors) to create and post tasks for others (workers or Turkers) to complete for compensation. Typically, completion of these tasks, called Human Intelligence Tests (HITs), offer monetary compensation sent directly to the Turker's account. The amount compensated is decided by the requestor and is often paid in U.S. cents, with Turker's salaries typically being around 1–3 US dollars per hour (Hara et al., 2019). Some Turkers also use the platform as a primary source of income, although those who do so are predominantly from India (Ross et al., 2010). Moreover, Turkers who use this platform as work are incentivized to complete many HITs to make a livable wage for full-time workers or additional income for part-time workers.

AMT has been used as a platform to collect data from a variety of disciplines and its usage has increased over time. In October 2022, a search was conducted on the terms "mechanical turk" or "AMT" across the databases Academic Search Ultimate, Medline, PsycInfo, and SocINDEX to demonstrate the multi-discipline usage of AMT samples. The only filter applied to the search was that the article was peer-reviewed (and there were no duplicates). The first ones found were in 2010 and the number of works found from 2010–2021 can be seen in Table 1 which shows a consistent increase over time.

**Table 1**

*Number of Publications Using AMT by Year*

| Year of Publication | Number of Publications |
|---|---|
| 2010 | 2 |
| 2011 | 14 |
| 2012 | 24 |
| 2013 | 47 |
| 2014 | 130 |
| 2015 | 274 |
| 2016 | 402 |
| 2017 | 490 |
| 2018 | 592 |
| 2019 | 715 |
| 2020 | 901 |
| 2021 | 1118 |

The use of AMT is not restricted to one academic area as it is used across a vast array of fields. Some recent examples include addictions (Mellis & Bickel, 2020), advertising (Connors et al., 2020), criminal justice (Fissel et al., 2021), geography (Kruse et al., 2021),

linguistics (Ciancia & Gallo, 2021), management (Brown et al., 2021), medicine (Lee et al., 2023), pharmaceutical sciences (Lin et al., 2021), political science (Blankenship et al., 2021), psychology (Ratcliff & Hendrickson, 2021), public health (Stevens et al., 2021), and sociology (Wilbur et al., 2021). Clearly, many fields have embraced the use of AMT as a data collection platform as it does offer some obvious benefits.

One benefit of AMT is the sheer reach and diversity of populations that use it, allowing for a greater diversity in participants available to researchers. Due to the popularity of purchasing products on Amazon, many individuals know of and use it by association, therefore leading to a large pool of participants. Previous research represented a very homogeneous makeup of Turkers, demonstrating that most are from the United States and India (Ross et al., 2010). More recent research still indicates a similar trend (Difallah et al., 2018) in country of origin of Turkers. Surveys by Difallah et al. (2018) indicate there are as many as 100,000 Turk users, with 2,000 of them being active at any one time. Due to the ubiquity of Amazon, and the convenience of AMT, large samples are easily obtainable and some cross-cultural studies can be conducted. There is also a wider variety of ages represented by Turkers compared to convenience samples, such as students in college (Ross et al., 2010). In addition, there is a fairly low cost of compensation that is paid to participants compared to other competitors (e.g., Prolific). Another benefit that researchers gain from using AMT that is paramount is the convenience. Data from individuals from diverse backgrounds and geographic locations can also be collected extremely quickly using the AMT platform depending on several factors related to the type of survey or research being conducted. Additionally, not all Turkers are the same as there are different categories of workers on AMT.

## Master Turkers

There are subsets of Turkers that may provide higher quality data and might be more sought out by researchers. Peer et al. (2014) found that the reputation of Turkers was significantly related to the quality of data. In the AMT space, reputation is measured by approval rating which is accomplished by requestors rating the Turkers on their performance on the HITs that they completed. Peer et al. (2014) reported that high reputation Turkers produced better data and using only high approval rated Turkers might be a viable strategy for improving overall data quality. Amazon does have functionality for selecting Turkers who have high approval and providing a different categorization for them. Turkers who complete many HITs can occasionally be "promoted" to an elevated status referred to as Master Turker. The Master Turker status is assigned to Turkers who complete various HITs and are consistently rated positively by requestors (Amazon, 2018). There is not currently a system to apply to be a Master Turker, as algorithms (not publicly available) are used to calculate some metric of number of HITs completed and ratings, leaving the actual qualifications somewhat nebulous.

With the status of Master Turkers seemingly seeking to rectify the problem of low-quality data, one would be brought to think that an overwhelming majority of researchers would use them as their sample population. However, there are several reasons why one may not do this, the first of which being the extra fee for using exclusively Master Turkers. Amazon charges an additional 5% for using the Master Turker qualification when selecting who the requestor's HIT is shown to (Amazon, 2018). Rouse (2020) completed two studies to evaluate Amazon's (2018) claim that Master Turkers provide higher quality data and are therefore worth the additional 5% premium. A series of two studies were conducted; the first experiment showed no difference between Masters and non-Masters on a personality assessment, whereas the second study used a 1-tailed test to determine if Masters produced higher reliability estimates on a cognitive ability test, which was not supported. In fact, if a 2-tailed test was used it would have shown a significant pattern in the opposite direction. The latter finding calls into question the claim from Amazon that Master Turkers should be compensated more because they provide higher quality work. The question to be addressed is if Master Turkers provide a different quality of data.

One explanation of differential quality of data among Turkers may be due to a subsample of Master Turkers who complete a large number of HITs as they optimize ways to complete tasks. Harms and DeSimone (2015) found that samples of these workers contribute a disproportionate amount of all HITs completed on AMT. Chandler et al. (2014) found that the top 1% of Turkers completed 11% of all HITs on the platform. This may be detrimental due to the increased likelihood of previous exposure to many survey and experiment paradigms as those Turkers become privy to them. Ford (2017) referred to *speeders* and *cheaters* which are Turkers who are incentivized to maximize the amount of money they can earn by speeding or skipping through HITs at a fast rate in order complete the task(s) and earn the reward while not providing accurate data. Harms and DeSimone (2015) described *Superturkers* as a group who spend an inordinate amount of time on AMT in an attempt to optimize their daily HIT completion rates. Thus, if participants are trying to complete a high number of HITs, the issue of data quality that should be considered among Turkers is attention (Buhrmester et al., 2018).

With these issues compounding over time, Amazon may experience less requestors. As an example, Chmielewski and Kucker (2020) performed a study that examined Turker performance on tasks over time which indicated there may be an AMT crisis in relation to data quality. They reported a pattern of failing response validity indicators, worse psychometric properties, and the inability to replicate well established findings over time. Their interpretation was that data quality was decreasing over the timeframe in which the study was conducted which included four rounds of data collection. Furthermore, some journals, editors, and reviewers have rejected manuscripts on the basis of using an AMT sample regardless of the study design and outcome (e.g., Landers & Behrend, 2015; Walter et al., 2019). Concerns over the frequency in which Turkers are exposed to certain

experimental paradigms, motivation to achieve compensation, selection bias (Landers & Behrend, 2015) and concerns over the measurement properties and characteristics of online samples like this (Walter et al., 2019) are some of the reasons why studies using populations like the Turk have been rejected. Amazon's policies and suggestions regarding requestors and Turkers clearly prioritize quantity over quality and it will be interesting to track the trajectory of research conducted using AMT samples.

Data quality is a composite of several facets relating to the collection of data, specifically, accuracy, which can be conceptualized as avoiding errors while collecting the data (Herrera & Kapur, 2007). For this study we will be analyzing this concept by comparing the difference in the act of straight lining (selecting the same answer for the entirety of a scale), which would indicate that the person likely isn't accurately reporting their scores, completion time, and the Cronbach's alpha reliability of differently coded scales in a study design similar to Rouse (2020). Rouse (2020) based this analysis design on a comparison of an in person sample lab sample and an MTurk sample conducted by Johnson and Borden (2012). This data quality is analyzed in the context of psychology survey tasks as this is a very common usage for the MTurk population.

The number of studies using AMT has continued to grow over time (see Table 1), yet there have been some more recent studies (e.g., Aguinis et al., 2021; Rouse, 2020) that have identified some red flags regarding data quality. The present study sought to further explore the relationship between data quality and Master Turker status to see if Amazon's algorithm of categorizing Master Turkers is related to higher quality data. Specifically, the question the study sought to investigate was if there are differences between Master Turkers and non-Master Turkers on data quality measures such as Cronbach's alpha reliability scores on frequently used survey instruments, completion time, and poor survey taking behavior such as straight lining through the survey. This was examined by administering two commonly used scales; one scale had 10-items that were anchored similarly, whereas the other scale had half of the items reverse coded.

# Method

## Participants

Demographic data was not collected for this survey; rather, information on participants' AMT-related behaviors were of interest in describing the sample. Our sample size was calculated based on Rouse (2020) which used Bonett's (2003) recommendation for testing the significance between two reliability estimates with a one tailed significance of .05 with a power of .80 and range of reliability estimates from .70 to .85. We decided to oversample to further increase statistical power. Overall, 320 participants took the survey, and 309 were analyzed after removing incomplete data. Table 2 provides the relevant characteristics of the Master Turkers while Table 3 provides data for non-Master

Turkers. In summary, 50% of participants self-reported being Master Turkers and 49% reported they were not. At the time of writing, there is no option to exclude Masters from the participant pool on MTurk, only to exclude non-Masters. English was the primary language by all but one participant. Approximately half of participants were Turkers for less than one year. Three out of four participants reported being full-time Turkers and over 40% reported completing more than 40 HITs per week.

**Table 2**

*Demographics of Master Turkers*

| Category | *N* | % |
|---|---|---|
| **Length of Turk usage** | | |
| < 6 months | 30 | 19.4 |
| 6–12 months | 45 | 29.0 |
| 1–5 years | 73 | 47.1 |
| > 5 years | 7 | 4.5 |
| **Full or part time** | | |
| Full time | 129 | 86.6 |
| Part time | 20 | 13.4 |
| No response | 6 | 4.5 |
| **Number of weekly Turk tasks completed** | | |
| 1–5 | 12 | 7.7 |
| 6–10 | 20 | 12.9 |
| 11–20 | 28 | 18.1 |
| 21–40 | 45 | 29.0 |
| 41–100 | 34 | 21.9 |
| > 100 | 16 | 10.3 |
| **First language** | | |
| English | 149 | 96.1 |
| Non-English | 5 | 3.2 |
| No response | 1 | 0.6 |

**Table 3**

*Demographics of Non-Master Turkers*

| Category | *N* | % |
|---|---|---|
| **Length of Turk usage** | | |
| < 6 months | 29 | 19.2 |
| 6–12 months | 48 | 31.8 |
| 1–5 years | 35 | 23.2 |

| Category | *N* | % |
|---|---|---|
| > 5 years | 39 | 25.8 |
| **Full or part time** | | |
| Full time | 104 | 68.9 |
| Part time | 47 | 31.1 |
| **Number of weekly Turk tasks completed** | | |
| 1–5 | 4 | 2.6 |
| 6–10 | 15 | 9.9 |
| 11–20 | 20 | 13.2 |
| 21–40 | 34 | 22.5 |
| 41–100 | 19 | 12.6 |
| > 100 | 59 | 39.1 |
| **Language** | | |
| English | 150 | 99.3 |
| Non-English | 1 | 0.7 |

## Instruments

### Rosenberg Self Esteem Scale

A 10-item scale that measures self-esteem by assessing positive and negative feelings about the self (Rosenberg, 1965). Items on the scale use a 4-point Likert scale format ranging from strongly agree to strongly disagree. Scores on the scale range from 10–40, with higher scores indicating higher self-esteem. One notable aspect of this scale is that several of the items are reverse coded (2, 5, 6, 8, 9), where a strongly disagree will indicate a higher self-esteem. The Rosenberg Self-Esteem Scale is a widely used and validated scale (Gray-Little et al., 1997). This scale was selected because of its widespread use and solid validation and is also relatively brief. Many analyses have been performed on the Rosenberg self-esteem scale which consistently provides respectable psychometric properties (e.g., Schmitt & Allik, 2005; Sinclair et al., 2010).

### PANAS Scale – Positive

The 10 positive items from the Positive and Negative Affect Scale were used to determine participants feelings of positive affect (Watson et al., 1988). PANAS uses a 5-point Likert scale ranging from 1 (very slightly or not at all) to 5 (extremely), and total scores can range from 10–50, with higher scores indicating higher amounts of positive affect. This scale was selected for its brief nature, as well as having 10-items of non-reverse coded items which is a reasonable comparison to the 10-item reverse coded self-esteem scale. Similarly, many analyses have consistently shown favorable psychometric properties of the scale across a wide range of samples (e.g., Crawford & Henry, 2004).

## Procedure

A survey was created using the software Qualtrics to be distributed through Amazon's AMT platform. The survey consisted of AMT use related questions, the two scales provided in a counter balanced manner, and some exploratory questions aimed to serve as the basis for a future study. Pilot testing of the survey consisted of 13 individuals and had a range of 4 minutes 17 seconds and 34 minutes 56 seconds, with an average completion time of 9 minutes and 56 seconds. With upper outliers removed, the average completion time of the pilot study was 5 minutes 31 seconds. Participants were awarded 0.12 USD upon completion of the survey as a result of the pilot study taking just over 5 minutes to complete. It was reasoned that 0.10 USD would be too little for over 5 minutes, so 0.12 USD was selected to reflect better value for the participant's time. The survey was limited to AMT users from the United States and was written in English. Participants had a completion time ranging from 48 seconds to 89 minutes 6 seconds. Average completion time for the survey was 4 minutes and 16 seconds. The survey purposely did not ask any personal demographic questions, and instead focused on questions related to AMT usage and behavior. The study was fielded for 8 hours from approximately 9 am to 5 pm EST. The primary goal of this study was to compare data quality for Master Turkers and non-Master Turkers. This was primarily assessed by comparing the Cronbach's alpha scores of these two groups on two scales, one reverse coded and one not. Cronbach's alphas were compared using the methodology developed for an online research environment by Diedenhofen and Musch (2016) which was based on previous methodology developed in Feldt et al. (1987). This method employs the use of a chi-square test to compare Cronbach's alpha scores. This differs from the typical benchmark comparison to .7 and allows for the comparison of two Cronbach's alphas relative to each other, based on sample size and number of items on the scale you are measuring the Cronbach's alpha of.

# Results

To asses completion time an independent sample *t*-test was conducted and found Masters ($M$ = 280, $SD$ = 497) were not significantly different, $t(207.1) = 1.54$, $p = .126$, than non-Masters ($M$ = 213, $SD$ = 207) on the average amount of time it took them to complete the survey. A non-parametric test was also used to evaluate completion time, and it was found there was also no significant difference in completion time based on Master status, $H(1) = .576$, $p = .448$. A chi-squared test determined there was no significant difference between Masters and non-Masters on frequency of straight lining behavior, $\chi^2(2, N = 306) = 8.07$, $p = .045$. Table 4 contains data on the frequency of straight lining. A one-way MANOVA was conducted to examine the difference in mean scores between Master and non-Master Turkers. There were two dependent variables: score on the Rosenberg Self

Esteem Scale and score on the positive PANAS scale. There was a significant main effect of whether someone was a Master Turker or not on the scores that participants obtained on the scales, $F(1, 304) = 8.58$, $p < .001$. Individual ANOVAs found a significant level of difference on the Rosenberg scale, $F(1, 304) = 7.03$, $p = .008$, such that Masters, $M = 24.36$, $SD = 3.64$, were higher than non-Masters, $M = 23.07$, $SD = 4.84$, as well as on the PANAS positive, $F(1, 304) = 7.99$, $p = .005$, where again Masters, $M = 39.54$, $SD = 6.29$, were higher than non-Masters, $M = 37.22$, $SD = 8.00$.

**Table 4**

*Straight Lining Behavior by Turker Status and Scale*

| Variable | No Straight Lining | PANAS | Rosenberg | Both | Total |
|---|---|---|---|---|---|
| Master | 137 | 5 | 9 | 4 | 155 |
| Non-Master | 126 | 6 | 19 | 0 | 151 |
| Total | 263 | 11 | 28 | 4 | 306 |

The Rosenberg scale demonstrated acceptable internal consistency for non-Master Turkers ($\alpha = .76$), and unacceptable consistency for Master Turkers ($\alpha = .34$). A Chi-squared test was used to determine whether these values differ significantly, $\chi^2(1, N = 304) = 30.05$, $p < .001$. The PANAS scale demonstrated acceptable internal consistency for Master Turkers, $\alpha = .73$, as well as for non-Masters, $\alpha = .82$. A Chi-squared test determined that these values differ significantly, $\chi^2(1, N = 304) = 5.00$, $p = .025$.

## Discussion

The data in this study revealed some surprising findings surrounding the performance of Master Turkers on completing commonly used instruments. Master Turkers had significantly less reliable data than what was provided by general Turker samples. This runs contrary to Amazon's (2018) claim that Master Turkers provide higher quality data. It should also be mentioned that the general Turker sample yielded reliability coefficients within the range that has been consistently reported (e.g., Crawford & Henry, 2004; Schmitt & Allik, 2005). Because of the premium associated with the use of Master Turkers and the seemingly worse data quality associated with them, the findings of this study suggest it may not be worthwhile to limit surveys to only using Master Turkers in studies that use instruments similar to the ones used in this study. In fact, the results suggest that the general Turker population provides significantly higher quality data for the two short instruments that were used in the study.

The design of the present study was to intentionally compare two instruments with the same number of items, differing on the basis of being reverse coded or not. Cron-

bachs alpha scores on the PANAS are frequently around the high .80s (Carvalho et al., 2013; Serafini et al., 2016; von Humboldt et al., 2017; Watson et al., 1988), of note the Master Turker population, $\alpha = .73$, had a significantly lower Cronbach's alpha score than the non-Masters, $\alpha = .82$, however both are still in the range that is conventionally considered acceptable. For the Rosenberg scale, a study across 53 different countries found an average Cronbach's alpha score of .81 (Schmitt & Allik, 2005). This is higher than both masters (.34) as well as non-Masters (.76). Non-Masters, however, have an acceptable Cronbach's alpha, whereas the Masters have an alpha that is considered to be far below what is acceptable. Clearly, Master Turkers completed the scales differently than the non-Master Turkers, particularly the Rosenberg with the reverse coded items. The low reliability for the Rosenberg stands in stark contrast to its widely used and accepted nature (Schmitt & Allik, 2005), and serves as a good indicator that the Master Turker population did not provide high quality data. Master Turkers are paid more than non-Master Turkers and have more experience, so it is important to consider factors that might be associated with their poor performance on these tasks.

The observed low level of reliability among Master Turkers may have a number of possible reasons to explain the findings. One consideration is that many Turkers use AMT as a full-time job (Ross et al., 2010) which was found in our study as roughly three out of four respondents indicated they did AMT full-time. Thus, it is within a full-time Turker's best interest to complete as many HITs as possible to maximize the amount of money they make in a given period of time. If a Turker is more motivated to complete a high quantity of HITs this would lead to justifying cheating, and speeding as Ford (2017) names them, to achieve this quantity. To compound this issue, if the requestor is conducting research alone with high quantities of participants, it can be challenging for them to validate and identify every individual participant and approve their response in the given timeframe. Additionally, Amazon (2018) advises the requestor to *not* reject work often, and that it is inappropriate to penalize a worker because of unclear instructions that the requestor provides. The reasoning provided is that Amazon is aware of how important Turker approval ratings are and will actively avoid requestors who are seen as harsh or unfair. With these statements officially posted on Amazon's approval guide for requestors, it encourages requestors to be lenient and provides an incentive system for Turkers to do as many HITs as possible within a given timeframe. This is problematic for several reasons including that it is encouraging requestors not to reject responses, because it will lead to them not having any Turkers willing to complete their HITs. This creates a cycle for these high-volume and efficient Turkers to have highly positive approval ratings, allowing them to be eligible for Master status and therefore higher pay. For Turkers who do this as a full-time job, this is the end goal, but for researchers this is a problem as it can be argued that AMT's policy suggestions do not optimize data quality.

Another factor that could be contributing to improving optimization of efficiency among frequent Turkers is their familiarity with common psychological scales and research paradigms. Due to the high volume of HITs that these high-volume Turkers complete, there is a modest likelihood that they have been exposed to a wide variety of scales and psychological paradigms. Because of this non-naiveté, these Turkers will complete these tasks much more quickly, and in a way that they believe is expected of them (Chandler et al., 2014). This is also another unfortunate downside of Turkers who complete a high number of HITs, because they have often been exposed to these scales and tests multiple times, they may be familiar with the measurement that is trying to be assessed and may answer in ways that they believe the scale or paradigm is supposed to be answered instead of answering truthfully.

This, however, does not indicate that Master Turkers are faster at completing individual surveys, as Masters and non-Masters were found not to differ significantly on the time it took them to complete the survey. This is interesting to note as previous research (Ford, 2017; Harms & DeSimone, 2015) seemed to indicate that those who achieve Master Turker status would be faster than non-Masters. The opposite is found in the present data, indicating on average, Masters took slightly more time taking a mean of 280 seconds to complete the survey, compared to non-Masters taking a mean of 213 seconds. However, the masters varied more as a population, indicated by the larger standard deviation.

A third possible explanation related to the prior two possible issues is that of attention. Attentiveness can vary throughout all research populations including AMT, however, in an online space, there exists many more distractors than in a controlled laboratory or classroom setting and the rapid speed at which MTurk users complete HITs to maximize monetary output may negatively impacts attention (Aguinis et al., 2021). Chandler et al. (2014) found that most Turkers reported that they completed HITs alone in their own home, they also reported doing other activities simultaneously such as watching television, listening to music, or instant messaging/texting. Thus, it is important to assess Turkers level of attention. One mechanism of doing so is by using some type of validity checks within the study. Oppenheimer et al. (2009) explored a possible solution to this lack of attention in general research populations by employing an instructional manipulation check (IMC), which mimics the format and length of other survey questions, but instead seeks to assess whether the participant is reading and interpreting a question. The idea behind the IMC is for the participant to fully read the question and ignore the response pattern that is typical for the rest of the survey. Throughout the study it was discovered that IMCs failure rate depended on the further context of the survey, including when the IMC was presented to the participants. It was also noted that even if the IMC question format did not fit with the context of the question asked, a small percentage of participants still failed the IMC. One such strategy is removing those who fail the IMC altogether to increase statistical power. Another

strategy to increase statistical power is by forcing participants to pay more attention by prompting them with the IMC, thereby priming them to read more throughout the survey.

A recent investigation (Aguinis et al., 2021) found that 15% of Turkers failed attention checks and were also likely to exhibit a myriad of behaviors that indicate a lack of attention. Interestingly, Lovett et al. (2018) surveyed Master Turkers, of which 70% believed their data to be of very high quality whereas the other 30% indicated it was high quality. In addition, the study had a qualitative component and Master Turkers reported that the factors that were related to high quality data were: higher compensation based on time, attention checks used in the HIT, more experience, higher reputation, multiple choice over writing, clear directions, and clean formatting. Unfortunately, the present study did not support the views reported by Lovett et al. (2018).

In contrast, a different pattern may exist for non-Master Turkers. Specifically, those Turkers may be working toward a Master Turker designation. The best way to achieve that status is to complete many HITs, but to do so well. Although AMT does not provide the algorithm or criteria for achieving Master Turker status, it is a reasonable assumption that part of the calculus involves some combination of number and accepted HITs. In other words, they have to complete quite a few tasks and receive strong ratings over some period of time. Achieving a high rating will most likely be related to successfully completing the tasks (i.e., HITs) and following the directions carefully.

This creates an interesting issue for researchers as well as for Amazon. For researchers, results from the present study might serve as a caution with regards to instruments to use (or not use) if using Master Turkers. It appears as though non-Master Turkers performed as samples drawn from other populations. Consider the possibility of a researcher limiting the respondents to Master Turkers which may result in paying more for lower quality data. Then, researchers may be faced with the daunting task of determining "which" data to keep.

A recent paper (Aguinis et al., 2021) provided a thorough summary of the benefits and validity threats associated with using Turkers as participants in a study. The authors provided a list of four benefits, ten threats as well as ten steps to consider when conducting a study on AMT. Of particular relevance to the present study was step eight which was labeled "screening data" which of course, is important in all studies. Aguinis et al. (2021) mentioned particular concerns in data screening involved BOTS, high attrition, and inattention with the possible remedies of attention checks, checking response times, estimating the number of useable responses prior to the study so that one oversamples knowing some data will be deleted, and examining response patterns. Data may need to be deleted, which then the findings (or lack thereof) may come into question. It is true that this dilemma is present in most types of research when data is collected from humans, but the AMT platform has additional nuanced issues to deal with and resolving

some of the relevant issues may become more pronounced over time as technology advances in this domain.

The MTurk population could also potentially be less representative than it appears based on demographic information. Studies have demonstrated that the average Turker is not very representative in age and may in fact be rather unusual in the level of educational attainment as well (Difallah et al., 2018; Redmiles et al., 2019; Ross et al., 2010). It is suggested to counteract this by being more selective with the options that MTurk provides, additional costs will be incurred but the data will be more representative (Zack et al., 2019).

Although this study had interesting findings, it does have limitations to consider. First, one limitation the present study faced was the demographic which consisted of only Turkers who indicated they lived in the United States, which may or may not be true. Much of the data surrounding AMT usage indicates the stark differences in the American and Indian populations that predominantly make up the userbase of the Turk (Ross et al., 2010). A second limitation involves the scales that were used. Both scales were brief 10-item scales that are well known. And, although that is a useful comparison, many research studies use scales with many more items which may provide different results. Both scales measured what might be considered dimensions of personality so scales tapping other domains would also be of interest. Third, minimal screening was done to recruit participants in AMT. It may be possible that more rigorous screening methods would result in different outcomes.

## Conclusions

The results of the present study provide a cautionary tale for potential requestors using AMT which is to prepare to screen data if Master Turkers are completing instruments with reverse coded items. This also highlights an issue that Amazon would benefit from addressing, as potential customers may be disincentivized to post their surveys on AMT due to questionable reliability from Masters at a higher price than non-Masters. It may be the case that this platform may require more methodological scrutiny and rigor than other data collection methodologies, particularly when using samples that may be somewhat risky with regards to the samples. When feasible, it may be worthwhile to conduct a form of replication or some comparison to a sample obtained from a different platform or setting. Certainly, AMT is not going away anytime soon but there are clear methodological issues to consider when using it for collecting data.

**Funding:** The authors have no funding to report.

**Acknowledgments:** The authors have no additional (i.e., non-financial) support to report.

**Competing Interests:** The authors have declared that no competing interests exist.

**Ethics Statement:** This study was deemed exempt for its use with human subjects and approved by the Shippensburg University IRB Human Subjects Committee (approval no. 2901) on May 04, 2022.

**Data Availability:** The authors confirm that the data supporting the findings of this study are available within the Supplementary Materials (see Trenge & Griffith, 2024a).

## Supplementary Materials

For this article, the following Supplementary Materials are available:

- Cleaned data with incomplete entries removed, both masters and non-masters mixed together (see Trenge & Griffith, 2024a)
- Exact item wording - Survey as it was presented to participants on AMT through qualtrics (see Trenge & Griffith, 2024b)

### Index of Supplementary Materials

Trenge, C., & Griffith, J. D. (2024a). *Supplementary materials to "Master Turkers: An assessment of data quality"* [Data]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.15023

Trenge, C., & Griffith, J. D. (2024b). *Supplementary materials to "Master Turkers: An assessment of data quality"* [Item wordings]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.15022

## References

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management, 47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Amazon. (2018). *FAQs.* Amazon Mechanical Turk. https://www.AMT.com/worker/help

Blankenship, B. T., Davis, T., Areguin, M. A., Savaş, Ö., Winter, D., & Stewart, A. J. (2021). Trust and tribulation: Racial identity centrality, institutional trust, and support for candidates in the 2020 US presidential election. *Analyses of Social Issues and Public Policy, 21*(1), 64–98. https://doi.org/10.1111/asap.12256

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

Brown, M. I., Grossenbacher, M. A., Martin-Raugh, M. P., Kochert, J., & Prewett, M. S. (2021). Can you crowdsource expertise? Comparing expert and crowd-based scoring keys for three

situational judgment tests. *International Journal of Selection and Assessment, 29*(3-4), 467–482. https://doi.org/10.1111/ijsa.12353

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. https://doi.org/10.1177/1745691617706516

Carvalho, H. W. D., Andreoli, S. B., Lara, D. R., Patrick, C. J., Quintana, M. I., Bressan, R. A., Melo, M. F. D., Mari, J. D. J., & Jorge, M. R. (2013). Structural validity and reliability of the Positive and Negative Affect Schedule (PANAS): Evidence from a large Brazilian community sample. *Revista Brasileira de Psiquiatria, 35*, 169–172. https://doi.org/10.1590/1516-4446-2012-0957

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological & Personality Science, 11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Ciancia, C., & Gallo, A. (2021). Linguistic fieldwork amid the COVID-19 pandemic: How social distancing is affecting data collection. *I-LanD Journal: Identity, Language and Diversity, 2*, 135–153. https://doi.org/10.26379/IL2021002_008

Connors, S., Spangenberg, K., Perkins, A. W., & Forehand, M. (2020). Crowdsourcing the implicit association test: Limitations and best practices. *Journal of Advertising, 49*(4), 495–503. https://doi.org/10.1080/00913367.2020.1806155

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 43*(3), 245–265.

De Man, J., Campbell, L., Tabana, H., & Wouters, E. (2021). The pandemic of online research in times of COVID-19. *BMJ Open, 11*(2), Article e043866. https://doi.org/10.1136/bmjopen-2020-043866

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science, 11*(1), 51–60. https://doi.org/10.1371/journal.pone.0121945

Difallah, D., Filatova, E., & Ipeirotis, P. (2018, February). Demographics and dynamics of Mechanical Turk workers. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 135–143). Association for Computing Machinery.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*(1), 93–103. https://doi.org/10.1177/014662168701100107

Fissel, E. R., Fisher, B. S., & Nedelec, J. L. (2021). Cyberstalking perpetration among young adults: An assessment of the effects of low self-control and moral disengagement. *Crime & Delinquency, 67*(12), 1935–1961. https://doi.org/10.1177/0011128721989079

Ford, J. B. (2017). Amazon's Mechanical Turk: A comment. *Journal of Advertising, 46*(1), 156–158. https://doi.org/10.1080/00913367.2016.1277380

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23*(5), 443–451. https://doi.org/10.1177/0146167297235001

Hara, K., Adams, A., Milland, K., Savage, S., Hanrahan, B. V., Bigham, J. P., & Callison-Burch, C. (2019). Worker demographics and earnings on Amazon Mechanical Turk: An exploratory analysis. In S. Brewster & G. Fitzpatrick (Eds.), *Extended abstracts of the 2019 chi conference on human factors in computing systems* (pp. 1–6). Assocation for Computing Machinery. https://doi.org/10.1145/3290607.3312970

Harms, P., & DeSimone, J. (2015). Caution! AMT workers ahead—Fines doubled. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*(2), 183–190. https://doi.org/10.1017/iop.2015.23

Herrera, Y. M., & Kapur, D. (2007). Improving data quality: Actors, incentives, and capabilities. *Political Analysis, 15*(4), 365–386. https://doi.org/10.1093/pan/mpm007

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*(4), 245–251. https://doi.org/10.1177/0098628312456615

Kruse, J., Kang, Y., Liu, Y. N., Zhang, F., & Gao, S. (2021). Places for play: Understanding human perception of playability in cities using street view images and deep learning. *Computers, Environment and Urban Systems, 90*, Article 101693. https://doi.org/10.1016/j.compenvurbsys.2021.101693

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*(2), 142–164. https://doi.org/10.1017/iop.2015.13

Lee, A. C. H., Madariaga, M. L. L., Liao, C., & Ferguson, M. K. (2023). Gender bias in judging frailty and fitness for lung surgery. *The Annals of Thoracic Surgery, 115*(2), 356–361. https://doi.org/10.1016/j.athoracsur.2021.11.013

Lin, S. Y., Thompson, H. J., Hart, L. A., Fu, M. C., & Demiris, G. (2021). Evaluation of pharmaceutical pictograms by older "Turkers": A cross-sectional crowdsourced study. *Research in Social and Administrative Pharmacy, 17*(6), 1079–1090. https://doi.org/10.1016/j.sapharm.2020.08.006

Lovett, M., Bajaba, S., Lovett, M., & Simmering, M. J. (2018). Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology, 67*(2), 339–366. https://doi.org/10.1111/apps.12124

Mellis, A. M., & Bickel, W. K. (2020). Mechanical Turk data collection in addiction research: Utility, concerns and best practices. *Addiction, 115*(10), 1960–1968. https://doi.org/10.1111/add.15032

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Ratcliff, R., & Hendrickson, A. T. (2021). Do data from mechanical Turk subjects replicate accuracy, response time, and diffusion modeling results? *Behavior Research Methods, 53*(6), 2302–2325. https://doi.org/10.3758/s13428-021-01573-x

Redmiles, E. M., Kross, S., & Mazurek, M. L. (2019, May). How well do my results generalize? comparing security and privacy survey results from MTurk, web, and telephone samples. In C. Kruegel & H. Schacham (Eds.), *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 1326–1343). IEEE.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press. https://doi.org/10.1126/science.148.3671.804

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010, April 10–15). Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI'10 extended abstracts on human factors in computing systems* (pp. 2863–2872). Association for Computing Machinery. https://doi.org/10.1145/1753846.1753873

Rouse, S. V. (2020). Reliability of AMT data from masters and workers. *Journal of Individual Differences, 41*(1), 30–36. https://doi.org/10.1027/1614-0001/a000300

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology, 89*(4), 623–642. https://doi.org/10.1037/0022-3514.89.4.623

Serafini, K., Malin-Mayor, B., Nich, C., Hunkele, K., & Carroll, K. M. (2016). Psychometric properties of the Positive and Negative Affect Schedule (PANAS) in a heterogeneous sample of substance users. *The American Journal of Drug and Alcohol Abuse, 42*(2), 203–212. https://doi.org/10.3109/00952990.2015.1133632

Sinclair, S. J., Blais, M. A., Gansler, D. A., Sandberg, E., Bistis, K., & LoCicero, A. (2010). Psychometric properties of the Rosenberg Self-Esteem Scale: Overall and across demographic groups living within the United States. *Evaluation & the Health Professions, 33*(1), 56–80. https://doi.org/10.1177/0163278709356187

Stevens, E. M., Villanti, A. C., Leshner, G., Wagener, T. L., Keller-Hamilton, B., & Mays, D. (2021). Integrating self-report and psychophysiological measures in waterpipe tobacco message testing: A novel application of Multi-Attribute Decision Modeling. *International Journal of Environmental Research and Public Health, 18*(22), Article 11814. https://doi.org/10.3390/ijerph182211814

von Humboldt, S., Monteiro, A., & Leal, I. (2017). Validation of the PANAS: A measure of positive and negative affect for use with cross-national older adults. *Review of European Studies, 9*(2), 10–19. https://doi.org/10.5539/res.v9n2p10

Walter, S. L., Seibert, S. E., Goering, D., & O'Boyle, E. H. (2019). A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology, 34*, 425–452. https://doi.org/10.1007/s10869-018-9552-y

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wilbur, D., Sheldon, K. M., & Cameron, G. (2021). Autonomy supportive and reactance supportive inoculations both boost resistance to propaganda, as mediated by state autonomy but not state reactance. *Social Influence, 16*(1), 1–11. https://doi.org/10.1080/15534510.2021.1908910

Zack, E. S., Kennedy, J., & Long, J. S. (2019). Can nonprobability samples be used for social science research? A cautionary tale. *Survey Research Methods, 13*(2), 215–227.