Graduate Student Journal of Psychology
Fall 2025 - Vol. 25

Caution when Crowdsourcing: Prolific as a Superior Platform Compared with MTurk

Daniel OConnell[1], Ashley Bautista[1], Clint Johnson[2], and Amanda Venta[1]
[1]*Department of Psychology, University of Houston, Houston, TX, USA*
[2]*Department of Psychology, Webster University, St. Louis, MO, USA*

Many researchers host surveys on online crowdsourcing platforms, such as Amazon's Mechanical Turk (MTurk) and Prolific. Online platforms promise a convenient way to meet sample size needs while drawing on diverse pools that might not otherwise participate in science. Yet, the quality of data obtained from these platforms is often questionable, so the collection must be closely monitored and reviewed. This study aimed to independently determine which crowdsourcing pool best serves researchers who plan to recruit for online surveys. To achieve this aim, we analyzed data from a recently completed study that drew participants from both MTurk and Prolific. We screened the collected data for both cost and quality, focusing on measures of attention, duration, and internal consistency. We found that only 9.89% of MTurk participants (N = 354) and 43.34% of Prolific participants (N = 345) produced high-quality data; Prolific also proved to be the more affordable option. Researchers considering these platforms for recruitment may weigh the evidence to make decisions when developing their own recruitment strategies. Finally, we highlight best practices for social scientists conducting online research, including additional survey and screening techniques.

Keywords: MTurk, Prolific, survey, crowdsourcing, data quality

Researchers have leveraged the internet for years, during which the use of crowdsourcing platforms has increased dramatically (Aguinis et al., 2021). Gosling and Mason (2015) extolled the use of the internet for research a decade ago, endorsing surveys conducted through crowdsourcing platforms to reduce costs and participant attrition. Moreover, it is just as easy now for researchers to collect survey data from undergraduates at their institution as it is to reach participants overseas, thereby reducing long-standing generalizability concerns (Best et al., 2001; Gosling & Mason, 2015). Online surveys bridge international borders; as of January 2023, 64.4% of the world's population was connected to the internet (Kemp, 2023). Yet, this approach to data collection has limitations, particularly regarding data quality. The aim of the present study was to investigate strengths of participant pools for social scientists to obtain high-quality data. To this end, we analyzed data from a completed study drawing participants from two major crowdsourcing platforms, MTurk and Prolific, and evaluated both data quality and cost.

History

Since 2005, Amazon's Mechanical Turk (MTurk) has promised to optimize efficiency, augment data collection, reduce researcher cost, and grant access to diverse participants (https://www.mturk.com/). Researchers (e.g., Aguinis et al., 2021; Smith et al., 2015) point to diverse participants, speed of data collection, and low cost as reasons for MTurk's widespread use. Yet, researchers have found that data quality and treatment of diverse populations on these platforms can suffer (e.g., Burnette et al., 2022). Indeed, Aguinis and colleagues (2021) highlighted ten areas in which MTurk is limited in its ability to collect high-quality data. These areas include participants lying about personal information (e.g., Webb & Tangney, 2022), lack of English fluency (e.g., Moss et al., 2021), and gathering data from professional survey takers (e.g., Cheung et al., 2017), all of which can reduce effect sizes (Chandler et al., 2015; Newman et al., 2021). The Webb and Tangney (2022) study serves as a provocative example of poor data quality collected via MTurk; just 14 of their 529 participants were reportedly "human beings" (p. 1). Webb and Tangney (2022) are not the only researchers to encounter of low-quality data from crowdsourcing platform participants (Bai, 2018; Simone, 2019; Stokel-Walker, 2018), and others have been critical of crowdsourcing from MTurk (Barends & Vries, 2019; Kennedy et al., 2020; Zack et al., 2019).

More recently, in 2014, Prolific came to the market, similarly guaranteeing a vetted, engaged, and more diverse participant pool from numerous countries with an emphasis on ethical pay (Peer et al., 2017; https://www.prolific.com/). Indeed, Prolific holds great potential to overtake MTurk as the optimal crowdsourcing platform (Palan & Schitter, 2018). Yet, direct comparisons between the two have yielded conflicting–and sometimes biased–results. For instance, Peer et al. (2017; 2022) portrayed Prolific as superior

in terms of participant attentiveness, comprehension, honesty, and reliability compared to MTurk and CloudResearch, but both studies were funded by the Prolific company. Conversely, Litman et al. (2021) responded in a paper sponsored by CloudResearch, a company that accesses MTurk participants and aims to improve upon Amazon's platform. Their results demonstrated superior data quality on MTurk when paired with the CloudResearch Toolkit. Given directly contrasting results, it is important to establish objective criteria to fairly compare MTurk and Prolific recruitment, including data quality and pricing.

Deciding between MTurk and Prolific

Cost

One basic and practical consideration is cost. Prolific mandates that researchers pay their participants an ethical wage (Newman et al., 2021), which is a minimum of $8 U.S. Dollars (USD)/hr (https://www.prolific.com/). Meanwhile, MTurkers earn a minimum pay of $0.01 USD per assignment (https://www.mturk.com/). Prolific charges a higher platform usage fee (25% base rate for academics) compared to MTurk (20% base rate). However, Prolific includes most participant specifiers (e.g., age or job) within their base cost, whereas MTurk requires researchers to pay additional fees. Therefore, when participants are compensated equally, MTurk is cheaper until specifiers are added for researchers recruiting a specific population (e.g., young adults). Since MTurk does not enforce a minimum wage, researchers may pay participants less. Crump et al. (2013) found that higher wages did not incentivize participants enough to provide higher-quality results, but it did result in lower dropout rates. Conversely, Litman et al. (2015) showed that monetary compensation is a primary driver for participation, tying data quality to compensation rates, thus directly contradicting findings by Crump and colleagues (2013).

Data Quality

Researchers also value the quality of their data when using crowdsourcing platforms. Data quality is a term comprising many factors (Douglas et al., 2023), operationalized herein as–on the high-quality end–higher rates of passing attention checks and task completion combined with lower rates of lying and deception. Multiple methods are often combined to make conclusions about data quality (Douglas et al., 2023). Some techniques–often used in tandem with others–include attention checks, survey duration, and internal consistency.

Most studies employ attention checks (Douglas et al., 2023). To evaluate attention, survey designers may ask participants to make a forced response, write an open-ended response demonstrating understanding, or perform unrelated tasks like math–though they vary in effectiveness (Abbey & Meloy, 2017). That said, checks like these are not without detractors. Hauser et al. (2018) demonstrated that manipulation checks can confound results, particularly when implemented incorrectly (e.g., attention question placement is not randomized).

Another indicator researchers can use to determine data quality is survey duration (Teitcher et al., 2015). By comparing individual participant survey durations to the average and pilot data, researchers can identify outlier durations (Matjašič et al., 2018). Participants who respond far too quickly can be identified as suspicious and of low quality (e.g., Goodrich et al., 2023).

A third way to evaluate data quality is through internal consistency (e.g., Douglas et al., 2023). One way to evaluate internal consistency is through Cronbach's alpha (α) (Cortina, 1993), as random responding contributes to low values (Fong et al., 2010). When values are low (see Cortina, 1993), especially compared to validated standards of a measure, researchers should be skeptical about the overall reliability of their data.

Previous MTurk and Prolific Comparisons

A few independent studies have been conducted to directly compare MTurk and Prolific, demonstrating Prolific as superior. Albert and Smilek (2023) observed greater disengagement among MTurk participants compared to those on Prolific, though they only included high-performing MTurk users. While using participants identified by the platforms as high-quality can be beneficial for getting attentive participants (Lu et al., 2022), it limits random selection and naive respondents–those who are unfamiliar with certain measures (Matthijsse et al., 2015). In another direct comparison, Douglas and colleagues (2023) conducted an independent analysis across MTurk, Prolific, CloudResearch, SONA, and Qualtrics with a well-powered 500 participants per pool. They concluded that Prolific and CloudResearch outperformed the other pools in terms of data quality, with no substantial differences between the two; both outperformed the unmodified

MTurk. They also highlight other relevant details, such as the price per quality participant, wherein Prolific ($1.90) was cheaper than CloudResearch ($2.00) and MTurk ($4.36). Yet, similar to Albert and Smilek (2023), Douglas et al. limited participants by only allowing those who had already completed 100 surveys, thereby rejecting naive participants. The authors further suggest that their results ought to be regularly replicated, as pool demographic compositions fluctuate over time. The present study builds on these prior works by directly comparing MTurk and Prolific without pre-established participant quality standards.

Current Study

The current study aimed to directly compare the quality and cost of data gathered from identical surveys posted on MTurk and Prolific. Most previous studies comparing MTurk have pre-screened for high-performing users, limiting naive participants. In contrast, our study compared recruitment between MTurk and Prolific with naive and non-naive participants, representing the recruitment efforts commonly seen in contemporary research. Secondary data analyses were conducted on data collected in a previously completed study. Ultimately, we sought to answer the research question: How do cost and data quality from participants recruited from MTurk and Prolific differ without pre-screens in place? This question was answered using a thorough screening process influenced by prior research crowdsourcing data quality.

Method

Participants

For the MTurk sample (n = 354), most participants were White (81.64%), heterosexual (82.49%), and male (61.30%), with an average age of 26.18 years (SD = 4.54). For the Prolific sample (n = 345), most participants were White (77.08%), heterosexual (63.03%), and female (67.05%), with an average age of 22.20 years (SD = 2.03).

Procedures

This study utilized data collected through Qualtrics on MTurk and Prolific platforms. The current study aimed to compare samples drawn from MTurk and Prolific for a broader study (see more https://osf.io/2n8ge), which was approved by the IRB at Saint Louis University. Two identical surveys–differing only by the inclusion of an ID number for MTurk participants–were launched on the morning of April 15, 2022. Inclusion criteria required participants to be English-speaking young adults aged 18-25 and living in the United States.

Participants were told that they would be providing the company 'OCEAN' with feedback on their newly developed dating application rooted in personality. In reality, the study aimed to investigate participant preferences for romantic partners based on perceived personality and weight. Nevertheless, we subjected participants to a realistic process of testing a dating app which allowed them to create an OCEAN profile, rate eight random profiles, provide qualitative and quantitative feedback on the "app," and rate 34 images as high or low in BMI/weight. All participants were compensated $2 USD for approximately 15 minutes of work ($8 USD/hr rate).

Measures

Big Five Factor Model of Personality

The Mini-IPIP (Donnellan et al., 2006), a measure based on the Big Five Factor Model of Personality (Goldberg, 1999), was included as a component of the profile-building process to assess personality and induce psychological realism. The Mini-IPIP has demonstrated strong validity and internal consistency as a personality inventory (Donnellan et al., 2006). This measure was used to compare internal consistency before and after the screening process through Cronbach's α levels.

Demographics

Demographics were gathered through the profile-building process. Data included age, race, gender identity, sexual orientation, height, weight, and marital status.

Duration Data

Total survey duration captured via Qualtrics was used to compare quality before and after the screening process. Based on pre-launch trials, participants were expected to take a maximum of 15 minutes to complete the survey.

Data Quality Screening Process

The data screening process was inspired by the Webb and Tangney (2022) study, wherein participants were screened out in a step-by-step process and removed from the participant pool. The calculations for the cost of each high-quality respondent were inspired by Douglas et al. (2023).

The sequential screening process consisted of four steps: (1) age, (2) self-reported seriousness, (3) sensible

open-ended responses, and (4) other sensible responses.

Participants outside the age inclusion criteria between 18-25, inclusive, were screened out. Then, the final question of the Qualtrics survey asked participants: "How seriously did you take this survey?" Responses ranged from 1-5, with 1 being "*not very serious*" and 5 being "*very serious*." Those who admitted to not taking the survey seriously were screened.

Two open-ended questions were analyzed to screen for unreasonable and duplicate responses. One of these questions asked participants to "Please briefly summarize the purpose of this survey," following the consent form (on a separate page). The second, towards the end, asked participants to "Provide any remaining thoughts on OCEAN here." Criteria for what was considered reasonable were developed *a priori* using manifest content analysis (Graneheim et al., 2017). Responses that were marked correct must have mentioned the words "develop," "personality," "test," "algorithm," or "dating app" and sufficiently explain the purpose of the study. Exactly identical response featured exactly the same words, spelling, capitalization, and punctuation were also screened out.

Two additional metrics were used to refine participant quality based on congruence. First, participants were asked to rate 34 images as high or low in BMI (https://osf.io/2n8ge). Two images (one male and one female) were presented twice to measure consistency. Second, participants who provided impossible heights and weights were screened.

Results

Data Quality on MTurk versus Prolific

Results from the screening process are summarized in Table 1 and explained below.

Age

Of the 354 MTurk and 345 Prolific participants, 125 of the MTurk participants reported an age outside the restricted age range on the survey. This left 229 (64.69%) MTurk and 345 (100%) Prolific participants for analysis, totaling 82.12% of the sample.

Seriousness

Two MTurk participants did not respond to this question, and one individual on the Prolific survey reported a rating of 2, meaning they did not take it seriously. This left the participant count at 227 (64.12%) for MTurk and 344 (99.71%) for Prolific, or 81.69% of the total.

Sensible Open-Ended Responses

About a third (113) of the remaining Prolific participants were removed for illogical or incorrect responses on one or both of the open-ended questions. An example of this type of response included, "the whole body of salt water that covers nearly three-fourths of the earth." As a result, 80 (22.60%) MTurk participants and 231 (66.96%) Prolific participants remained, or 44.35% of the total sample. Next, identical responses were removed. For example, the response "OCEAN developers to improve the algorithm of their new dating app." appeared three times on MTurk. This affected participants in both pools such that 64 (18.08%) MTurk and 229 (66.38%) Prolific participants, or 41.92% of the total, remained.

Other Sensible Responses

First, participants were screened for inconsistent responses to identical questions. Of the remaining participants, just 38 (10.73%) MTurk and 153 (44.35%) Prolific participants, or 27.32%, were consistent in rating both sets of images at this stage. Next, participants were screened for impossible heights and weights. This affected three participants on MTurk for entering: (1) 8 feet 8 inches while weighing 120 pounds, (2) a height of 1 foot 1 inch tall, and (3) a weight of 154324 pounds. After this step, 35 (9.89%) MTurk and 153 (44.35%) Prolific participants remained, representing 188 of the initial 699 (26.90%).

Internal Consistency

Table 2 compares internal consistency on the Mini-IPIP between MTurk and Prolific alongside the original psychometric study (Donnellan et al., 2006). Both the Prolific and MTurk α values improved substantially after screening. Although Prolific scores generally began higher, both the MTurk and Prolific pre-screen data would be considered unreliable (Cortina, 1993). Moreover, after screening, all of the α values were higher for Prolific except for Intellect/Imagination. As a result, the evidence would support the post-screen reliability in Prolific but not MTurk due to values below .70 (Cortina, 1993).

Duration

Total survey duration was used to compare quality before and after the participant screening process. Based on pre-launch trials, participants were expected to take up to 15 minutes to complete the survey. The times that participants took on MTurk before (*Mdn*

= 10 minutes and 20 seconds) and after (*Mdn* = 10 minutes and 18 seconds) screening were slightly longer than the times that participants took on Prolific before (*Mdn* = 8 minutes and 47 seconds) and after (*Mdn* = 9 minutes and 2 seconds) screening. Using 2 *SD*s from the mean in each sample as a metric to compare speed (Matjašič et al., 2018), no responses on either MTurk or Prolific were considered outliers in the "fast" direction. While a handful of slow outliers were present, this was not meaningful to this study, as participants had the freedom to open the survey and complete it the following day.

Cost

Prolific was cheaper based on the total cost compared with MTurk. Costs included the direct payment to participants, the base hosting fee paid to the platform, additional specifier fees, and taxes. A total of $1,155 was paid to MTurk, compared with $979 for Prolific, a difference of $176. The difference comes primarily from MTurk's "Premium Qualifications" fee, which cost $0.50 extra per participant to recruit only participants aged 18-25. The cost per high-quality participant was also calculated by dividing the total cost by the respective number of users who produced high-quality data (Douglas et al., 2023). Prolific ($6.40 per high-quality participant) was still cheaper than MTurk ($33 per high-quality participant).

Discussion

The aim of the current study was to compare the practical and data-driven differences between two popular participant pools, MTurk and Prolific, building upon work by Douglas et al. (2023), Webb and Tangney (2022), and others. Data analyzed in this study were drawn from a completed study conducted primarily to make conclusions about online dating behavior in young adults, with data collected across two crowdsourcing platforms: Amazon's Mechanical Turk and Prolific. This comparison sought to understand the cost and quality of data gathered across both platforms. Based on pricing and data quality–assessed through attention checks, duration, and internal consistency–Prolific proved to be the superior crowdsourcing platform compared to MTurk for these samples. Nonetheless, Prolific still demonstrated notable room for improvement within this sample.

In this study, only about a quarter of the sample produced high-quality data. Of the 188 that remained after screening, most (n = 153) came from Prolific, compared with MTurk (n = 35). Nearly 18% of MTurk participants fell outside the inclusionary age range–despite the added cost–an effect also observed by Webb and Tangney (2022). As a result, even the mean age (26.18) was outside of the inclusion criteria range (18-25). Internal consistency further supported Prolific; Cronbach's α values were higher for all factors except for Intellect/Imagination. Notably, the change in α values after screening demonstrates that participant exclusion based on data quality can alter conclusions, an idea supported by previous research (DeSimone & Harms, 2018). Finally, duration of the survey appeared equivalent for Prolific and MTurk.

Prolific also outperformed MTurk on cost. On an absolute basis, Prolific was cheaper ($979 USD) compared to MTurk ($1,155 USD) for gathering the same number of participants (n = 350). While compensation for the participants was held even ($2), the host fees and specifier charges led to the observed differences. A steep increase in cost may lead to a moral conundrum in which researchers may lower participant wages to afford the hosting of their survey. As a better alternative, we recommend opting for a cheaper platform, which depends on exclusion criteria (i.e., base rate and the need for specifiers). On a relative basis, Prolific was still the cheaper option. As determined through the cost per high-quality participant, MTurk participants necessitated $33 compared to $6.40 for Prolific participants. Effectively, we paid MTurk five times the U.S. dollar value for fewer "usable" participants. Based on this detailed comparison of the samples gathered, the authors perceive Prolific as the winner in this direct comparison between MTurk and Prolific.

Limitations & Future Directions

Several platform capabilities were not tested in the present study. This study was not longitudinal, so the tools that both companies offer for this type of research could not be compared as they have in other studies (e.g., Henderson et al., 2021; Kothe & Ling, 2019; Paas et al., 2018; Stoycheff, 2016). Additionally, this was an experimental psychology study that took around 15 minutes to complete. There is reason to believe that studies presented in different fields (e.g., Follmer et al., 2017; Reid et al., 2022; Wagner et al., 2021) and durations (e.g., Aguinis et al., 2021; Hamby & Taylor, 2016) may find different success with each platform.

Additionally, it is difficult to determine the source of low-quality data. It is quite common to read papers that describe the data spoilers as "bots" (e.g., Goodrich et al., 2023; Stokel-Walker, 2018; Webb & Tangney, 2022). However, deeper dives suggest that international participants, not "bots" or computer programs, are a primary source of lower data quality (Moss et al., 2021). International participants are often excluded, so they may lie about demographic information (e.g., native language and current location), which can confound results (Dennis et al., 2020). It is recommended that further research be conducted on these topics. Moreover, a reproduction of this study is warranted to evaluate ever-changing pools.

Recommendations for Researchers

As researchers develop increasingly sophisticated methods to detect low-quality data or robots, participants and programmers evolve strategies to evade detection. While there is no perfect solution, steps can be taken by researchers and crowdsourcing companies to improve the science generated on these platforms by filling their online surveys with relevant attention checks, participant verifiers, and logic.

Goodrich et al. (2023) recommend considering embedded survey components, including CAPTCHA, honeypot questions, and institutional knowledge checks to improve participant screeners. CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) vary in form, including checking the "I'm not a robot" box, selecting all of the pieces of the stoplight in a given picture, or decoding distorted letters (Moradi & Keyvanpour, 2015). Honeypot questions are ones embedded and hidden in surveys, making them invisible to human survey takers but visible to robots (Goodrich et al., 2023). If one is answered, you have good evidence that your survey was answered by an actual robot. Finally, institutional knowledge can be checked in a similar way to the aforementioned logic check implemented in the present study. Goodrich and colleagues (2023) suggest a question about the participant's zip code and then a follow-up about a nearby landmark, such as the closest university.

IP addresses can also be used to vet participants who have signed up more than once in one location (Aguinis et al., 2021). Unfortunately, several drawbacks are present when collecting IP addresses. Anonymity is violated, prohibiting a guarantee of identity protection. Moreover, even if identifiable data are secured, as they should be, this check would not guarantee that the participant is only completing the survey once. Most survey takers know that they can use free VPNs (Virtual Private Networks) that allow them to appear, to internet service providers, as if they are in different places across the world (Dennis et al., 2020). This also may unfairly disqualify multiple individuals who use the same device to participate, such as public library computers or devices shared between family members.

Aguinis and colleagues (2021) recommend considering response speed and consistency in the process of screening participants. Apart from reviewing the entire survey time, which should fall around a certain predetermined duration based on trials, researchers can look at individual question response times. It is unlikely that participants could respond to certain questions in under a second (Wood et al., 2017) unless they are extremely familiar with a given measure or the objective is to respond rapidly. Therefore, tracking question response time, can alert researchers to suspicious data. Moreover, inattentive participants can be identified if they mark the same response several times in a row (e.g., "*strongly agree*" for all ten questions on a given measure; Aguinis et al., 2021). Several methods exist to analyze response patterns of this sort that may be used to flag bots (DeSimone & Harms, 2018; Dunn et al., 2018).

Finally, researchers should become aware of techniques not implemented in this study or discussed herein to identify participants who supply low-quality data, lie about answers, or submit multiple responses. Several researchers have done excellent work in compiling recommendations, which should be reviewed in tandem with reflection on this paper (Aguinis et al., 2021; Goodrich et al., 2023; Hunt & Scheetz, 2019; Hydock, 2018; Kennedy et al., 2020; Newman et al., 2021; Sauter et al., 2020; Stanton et al., 2022).

Conclusion

This study leveraged a screening process similar to Webb and Tangney (2022), with heavy influence from Douglas et al. (2023), to compare MTurk and Prolific recruitment potential based on data quality and cost. Based on these metrics, Prolific outperformed MTurk for recruitment. However, while Prolific outperformed MTurk on our survey, researchers with different protocols may observe different results. Most meaningful-

ly, researchers ought to critically evaluate the impact that using low-quality data in publications may have on societal outcomes for generations. As we found surprisingly few high-quality participants across both Prolific and MTurk, it is clear that improved survey methodologies are warranted regardless of platform. With this in mind, researchers should incorporate survey strategies demonstrated in this work as well as the highlighted best practices from other researchers.

References

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management, 53–56*, 63–70. https://doi.org/10.1016/j.jom.2017.06.001

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

Albert, D. A., & Smilek, D. (2023). Comparing attentional disengagement between Prolific and MTurk samples. *Scientific Reports, 13*(1), 20574. https://doi.org/10.1038/s41598-023-46048-5

Amazon Mechanical Turk. (n.d.). Retrieved March 30, 2024, from https://www.mturk.com/

Bai, H. (2018). Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates. *Maxhuibai.com*. https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random

Barends, A. J., & de Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences, 143*, 84–89. https://doi.org/10.1016/j.paid.2019.02.015

Best, S. J., Krueger, B., Hubbard, C., & Smith, A. (2001). An assessment of the generalizability of Internet surveys. *Social Science Computer Review, 19*(2), 131–145. https://doi.org/10.1177/089443930101900201

Burnette, C. B., Luzier, J. L., Bennett, B. L., Weisenmuller, C. M., Kerr, P., Martin, S., Keener, J., & Calderwood, L. (2022). Concerns and recommendations for using Amazon MTurk for eating disorder research. *International Journal of Eating Disorders, 55*(2), 263–272. https://doi.org/10.1002/eat.23614

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science, 26*(7), 1131–1139. https://doi.org/10.1177/0956797615585115

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology, 32*(4), 347–361. https://doi.org/10.1007/s10869-016-9458-5

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE, 8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting, 32*(1), 119–134. https://doi.org/10.2308/bria-18-044

DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology, 33*(5), 559–577. https://doi.org/10.1007/s10869-017-9514-9

Donnellan, M., Oswald, F., Baird, B., & Lucas, R. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18*, 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE, 18*(3), e0279720. https://doi.org/10.1371/journal.pone.0279720

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology, 33*(1), 105–121. https://doi.org/10.1007/s10869-016-9479-0

Follmer, D. J., Sperling, R. A., & Suen, H. K. (2017). The role of MTurk in education research: advantages, issues, and future directions. *Educational Researcher, 46*(6), 329–334. https://doi.

org/10.3102/0013189X17725519

Fong, D. Y., Ho, S. Y., & Lam, T. H. (2010). Evaluation of internal reliability in the presence of inconsistent responses. *Health and Quality of Life Outcomes, 8*, 27. https://doi.org/10.1186/1477-7525-8-27

Goodrich, B., Fenton, M., Penn, J., Bovay, J., & Mountain, T. (2023). Battling bots: Experiences and strategies to mitigate fraudulent responses in online surveys. Applied Economic *Perspectives and Policy, 45*(2), 762–784. https://doi.org/10.1002/aepp.13353

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Graneheim, U. H., Lindgren, B. M., & Lundman, B. (2017). Methodological challenges in qualitative content analysis: A discussion paper. *Nurse Education Today, 56*, 29-34. https://doi.org/10.1016/j.nedt.2017.06.002

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement, 76*(6), 912–932. https://doi.org/10.1177/0013164415627349

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00998

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Henderson, E. L., Simons, D. J., & Barr, D. J. (2021). The trajectory of truth: A longitudinal study of the illusory truth effect. *Journal of Cognition, 4*(1), 29. https://doi.org/10.5334/joc.161

Hunt, N. C., & Scheetz, A. M. (2019). Using MTurk to distribute a survey or experiment: methodological considerations. *Journal of Information Systems, 33*(1), 43–65. https://doi.org/10.2308/isys-52021

Hydock, C. (2018). Assessing and overcoming participant dishonesty in online data collection. *Behavior Research Methods, 50*(4), 1563–1567. https://doi.org/10.3758/s13428-017-0984-5

Kemp, S. (2023, January 26). *Digital 2023: Global overview report.* DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2023-global-overview-report

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods, 8*(4), 614–629. https://doi.org/10.1017/psrm.2020.6

Kothe, E. J., & Ling, M. (2019, September 6). Retention of participants recruited to a multi-year longitudinal study via Prolific. *PsyArXiv.* https://doi.org/10.31234/osf.io/5yv2u

Litman, L., Moss, A., Rosenzweig, C., & Robinson, J. (2021). Reply to MTurk, Prolific or panels? Choosing the right audience for online research. *SSRN*. https://doi.org/10.2139/ssrn.3775075

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods, 47*(2), 519–528. https://doi.org/10.3758/s13428-014-0483-x

Lu, L., Neale, N., Line, N. D., & Bonn, M. (2022). Improving data quality using Amazon Mechanical Turk through platform setup. *Cornell Hospitality Quarterly, 63*(2), 231–246. https://doi.org/10.1177/19389655211025475

Matjašič, M., Vehovar, V., & Manfreda, K. L. (2018). Web survey paradata on response time outliers: A systematic literature review. *Advances in Methodology and Statistics, 15*(1), 23-41. https://ibmi.mf.uni-lj.si/mz/2018/no-1/Matjasic2018.pdf

Matthijsse, S. M., de Leeuw, E. D., & Hox, J. J. (2015). Internet panels, professional respondents, and data quality. *Methodology, 11*(3), 81–88. https://doi.org/10.1027/1614-2241/a000094

Moradi, M., & Keyvanpour, M. (2015). CAPTCHA and its alternatives: A review. *Security and Communication Networks, 8*(12), 2135–2156. https://doi.org/10.1002/sec.1157

Moss, A. J., Rosenzweig, C., Jaffe, S. N., Gautam, R., Robinson, J., & Litman, L. (2021). Bots or inattentive humans? Identifying sources of low-quality data in online platforms. *PsyArXiv.* https://doi.org/10.31234/osf.io/wr8ds

Newman, A., Bavik, Y. L., Mount, M., & Shao, B. (2021). Data collection via online platforms: challenges and recommendations for future research. *Applied Psychology, 70*(3), 1380–1402. https://doi.org/10.1111/apps.12302

Paas, L. J., Dolnicar, S., & Karlsson, L. (2018). Instructional manipulation checks: A longitudinal analysis with implications for MTurk. *International Journal of Research in Marketing, 35*(2), 258–269. https://doi.org/10.1016/j.ijresmar.2018.01.003

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Prolific (n.d.). Retrieved March 30, 2024, from https://www.prolific.com/

Reid, B., Wagner, M., d'Amorim, M., & Treude, C. (2022). Software engineering user study recruitment on Prolific: An experience report. *ArXiv*. https://doi.org/10.48550/arXiv.2201.05348

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences, 10*(4), 251. https://doi.org/10.3390/brainsci10040251

Simone, M. (2019, November 21). Bots started sabotaging my online research. I fought back. *STAT*. https://www.statnews.com/2019/11/21/bots-started-sabotaging-my-online-research-i-fought-back/

Smith, N. A., Sabat, I. E., Martinez, L. R., Weaver, K., & Xu, S. (2015). A convenient solution: using MTurk to sample from hard-to-reach populations. *Industrial and Organizational Psychology, 8*(2), 220–228. https://doi.org/10.1017/iop.2015.29

Stanton, K., Carpenter, R. W., Nance, M., Sturgeon, T., & Villalongo Andino, M. (2022). A multisample demonstration of using the prolific platform for repeated assessment and psychometric substance use research. *Experimental and Clinical Psychopharmacology, 30*(4), 432–443. https://doi.org/10.1037/pha0000545

Stokel-Walker, C. (2018, August 10). Bots on Amazon's Mechanical Turk are ruining psychology studies. *New Scientist*. https://www.newscientist.com/article/2176436-bots-on-amazons-mechanical-turk-are-ruining-psychology-studies/

Stoycheff, E. (2016). Please participate in Part 2: Maximizing response rates in longitudinal MTurk designs. *Methodological Innovations, 9*. https://doi.org/10.1177/2059799116672879

Teitcher, J. E. F., Bockting, W. O., Bauermeister, J. A., Hoefer, C. J., Miner, M. H., & Klitzman, R. L. (2015). Detecting, preventing, and responding to "fraudsters" in Internet research: Ethics and tradeoffs. *Journal of Law, Medicine & Ethics, 43*(1), 116–133. https://doi.org/10.1111/jlme.12200

Wagner, A., Bakas, A., Kennison, S., & Chan-Tin, E. (2021). A comparison of SONA and MTurk for cybersecurity surveys. *Proceedings of the 2021 European Interdisciplinary Cybersecurity Conference*, 87–88. https://doi.org/10.1145/3487405.3487657

Webb, M. A., & Tangney, J. P. (2022). Too good to be true: Bots and bad data from Mechanical Turk. *Perspectives on Psychological Science. 19*(6), 887-890. https://doi.org/10.1177/17456916221120027

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science, 8*(4), 454–464. https://doi.org/10.1177/1948550617703168

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. https://doi.org/10.18148/srm/2019.v13i2.7262

Table 1

Summary of Results from the Screening Process

Screener Steps	MTurk (*n* = 354)	Prolific (*n* = 345)	Total (*n* = 699)
Age	229 (64.69%)	345 (100%)	574 (82.12%)
Seriousness	227 (64.12%)	344 (99.71%)	571 (81.69%)
Open-ended	64 (18.08%)	229 (66.38%)	293 (41.92%)
Other Sensible	35 (9.89%)	153 (44.35%)	188 (26.90%)
	Cost		
Total Cost	$1,155	$979	$2,134
Cost per high-quality participant	$33	$6.40	$11.35

Table 2

Reliability Metrics for the Validated Mini-IPIP (Donnellan et al., 2006), MTurk, and Prolific

	Mini-IPIP				MTurk			Prolific		
	α	Mean	*SD*		α	Mean	*SD*	α	Mean	*SD*
Extraversion	.77	3.28	.90	Before	.35	2.94	.76	.42	2.82	.26
				After	.80	2.67	.21	.86	2.72	.26
Agreeableness	.70	4.01	.69	Before	.29	3.26	.82	.51	4.02	.23
				After	.72	3.74	.36	.78	3.99	.21
Conscientiousness	.69	3.42	.78	Before	.26	3.13	.79	.38	3.58	.30
				After	.38	3.58	.24	.77	3.56	.34
Neuroticism	.68	2.54	.80	Before	.002	2.88	.83	.41	2.95	.38
				After	.57	2.78	.37	.78	2.91	.39
Intellect/ Imagination	.65	3.70	.73	Before	.52	2.98	.64	.37	4.00	.12
				After	.83	3.52	.05	.73	3.99	.12

Note. "Before" signifies the data prior to screening, and "After" signifies the data following screening.