Behavior Research Methods (2025) 57:340
https://doi.org/10.3758/s13428-025-02852-7

ORIGINAL MANUSCRIPT

Why you shouldn't trust data collected on MTurk

Cameron S. Kay[1,2]

Received: 15 April 2025 / Accepted: 19 September 2025

Abstract

Several prior studies have used advanced methodological techniques to demonstrate that there is an issue with the quality of data that can be collected on Amazon's Mechanical Turk (MTurk). The goal of the present project was to provide an accessible demonstration of this issue. We administered 27 semantic antonyms—pairs of items that assess clearly contradictory content (e.g., "I talk a lot" and "I rarely talk")—to samples drawn from Connect ($N_1 = 100$), Prolific ($N_2 = 100$), and MTurk ($N_3 = 400$; $N_4 = 600$). Despite most of these item pairs being negatively correlated on Connect and Prolific, over 96% were *positively* correlated on MTurk. This issue could not be remedied by screening the data using common attention check measures nor by recruiting only "high-productivity" and "high-reputation" participants. These findings provide clear evidence that data collected on MTurk simply cannot be trusted.

Keywords Amazon's Mechanical Turk · Data quality · Careless responding · Invalid responding · Survey design and methodology · Psychometrics

Introduction

There are two things to establish at the outset. The first is that Mechanical Turk (MTurk), the crowdsourcing platform introduced by Amazon in 2005, is incredibly popular. It has been used in countless psychology studies. Over 40% of the articles published in the *Journal of Personality and Social Psychology* and *Personality and Social Psychology Bulletin* in the first half of 2015 included at least one MTurk study, as did approximately 18% of the studies published in *Psychological Science* (Zhou & Fishbach, 2016; see also Goodman & Paolacci, 2017; Porter et al., 2019). The second thing to establish is that MTurk's popularity is not undeserved. Conducting studies on the internet carries numerous advantages (see Reips, 2000), and MTurk was one of the first platforms to make these advantages widely available to researchers. With MTurk, researchers (or at least researchers with funds) could collect data in a fraction of the time that was required using traditional methods; instead of taking several academic terms to collect data, data could be collected in an afternoon. Researchers could also move beyond the undergraduate student samples that dominated psychological research (Arnett, 2008; Thalmayer et al., 2021) to samples that were, at least in theory, more generalizable to the wider human experience (Berinsky et al., 2012).

Over the last several years, however, there has been growing concern about the quality of data that can be collected on MTurk (Douglas et al., 2023; Moss et al., 2021; Peer et al., 2022; Stagnaro et al., 2024; Webb & Tangney, 2022). Researchers have estimated the percentage of careless respondents on the platform to be somewhere in the range of 73.60% (Douglas et al., 2023) to 97.40% (Webb & Tangney, 2022; but, importantly, see Cuskley & Sulik, 2024), with approximately 40% of MTurk participants reporting that they lie when they respond to surveys on the platform (Stagnaro et al., 2024) and approximately 33% reporting that their responses may not be entirely accurate (Kay, 2024). Even the mere existence of Amazon's *Master Workers*—MTurk workers described as demonstrating "superior performance while completing thousands of tasks" (Amazon Web Services, 2024)—indicates that the average MTurk worker is likely not providing high-quality data.

The purpose of the present preregistered project was to provide a highly accessible demonstration of the data quality issues on MTurk. As elucidated above, there are several

✉ Cameron S. Kay
cameronstuartkay@gmail.com

[1] Environmental Social Sciences Department, Stanford University, 473 Via Ortega, Stanford, CA 94301, USA

[2] Psychology Department, Union College, 807 Union Street, Schenectady, NY 12308, USA

Published online: 10 November 2025

excellent studies that have used advanced statistical and methodological techniques to characterize the problem. Here, we wanted to provide a demonstration that makes the issue self-evident. To do so, we make use of what are called "semantic antonyms."

Semantic antonyms refer to any two items that are, on their face, contradictory (Goldberg & Kilkowski, 1985; see also Butcher et al., 2009). For example, the items "I talk a lot" and "I rarely talk" are semantic antonyms, as are "I like order" and "I crave chaos." Even without any domain knowledge, it should be clear that a participant should respond to these items in different ways. A participant who agrees that they talk a lot should disagree that they rarely talk; a participant who disagrees that they talk a lot should agree that they rarely talk. If a participant does not respond to these items in different ways, it is evidence that they are likely not paying attention to the content of the items they are responding to.

Particularly relevant to the present project, the correlation between semantic antonyms can be used as an index of the data quality on a data-collection platform. If one finds that pairs of these items are uncorrelated in a sufficiently sized sample, it indicates that enough participants are providing unrelated (and presumably careless) responses to obscure the true negative correlations between the items. Similarly, if one finds that pairs of these items are *positively* correlated on a data-collection platform, it indicates that there are enough participants providing similar (and, again, presumably careless) responses to overwhelm any true negative correlations.

How would careless respondents, who are presumably responding at random, ever result in a positive correlation between semantic antonyms? The short answer is that careless respondents do not always respond at random. In fact, humans are ill-equipped to produce random responses (e.g., Bakan, 1960; see Falk & Konold, 1997, for a review). In many cases, careless respondents simply respond by selecting the same response to multiple items in a row (i.e., straightlining; Johnson, 2005) or, if they are trying to avoid being flagged for careless responding, repeatedly cycling through a small set of responses (i.e., bandlining; Kay, 2024), such as bouncing back and forth between "Agree" and "Strongly agree" on the response scale. This can result in positive correlations between semantic antonyms since a participant's response to any one item will be similar to their response to every other item.

At least one prior study has indicated that there are enough participants engaged in this form of careless responding on MTurk to result in positive correlations between semantic antonyms. Specifically, in the process of validating two attention check measures, Kay (2024) administered three semantic antonyms to a sample of 562 MTurk participants. He found that "I am talkative" was positively correlated with "I don't tend to talk a lot" ($r = .32$); "I go through money quickly" was positively correlated with "I am good at saving my money" ($r = .31$); and "I do not sleep well" was positively correlated with "I sleep soundly" ($r = .54$).

The present project aims to extend these results by investigating whether this pattern holds for a larger set of semantic antonyms. It is possible (although not necessarily plausible) that the three semantic antonyms used by Kay (2024) were, in fact, semantic *synonyms*. This would mean that the positive correlations observed on MTurk were an indicator of *satisfactory* (or even *good*) data quality. To address this possibility, we increased the number of semantic antonyms considered to 27 and compared the correlations observed on MTurk to those observed on platforms generally known for having higher quality data (i.e., CloudResearch Connect; Prolific). If we find that the semantic antonyms are positively correlated on MTurk but negatively correlated on the other platforms, it would indicate there are enough participants engaged in careless responding on MTurk to produce positive correlations between items that are, by all accounts, contradictory.

While the idea that MTurk has a data quality issue is not new, finding a positive correlation between a large set of semantic antonyms on the platform would provide a clear demonstration of just how dire the situation is. Prior research has shown that careless responding can both attenuate (Credé, 2010; DeSimone et al., 2018; Hough et al., 1990; Oppenheimer et al., 2009) and inflate (Credé, 2010; Cornell et al., 2012; DeSimone et al., 2018; Holtzman & Donnellan, 2017; Huang et al., 2015; Zorowitz et al., 2023) observed effect sizes. The current study would be among the first to empirically demonstrate that it can also *reverse* the direction of observed effects and that the conditions to produce such spurious results are already present on MTurk.

Current project

The current project includes four preregistered studies. Given the similarity between the studies, varying only in the online data-collection platform used, we do not present them *in series* (i.e., one after the other). Instead, we present them *in parallel* (i.e., simultaneously). For simplicity, we refer to them simply as the Connect study, Prolific study, open MTurk study, and qualified MTurk study.[1]

[1] The Connect study was conducted first, followed by the open MTurk study, the Prolific study, and, finally, the qualified MTurk study.

We conducted the Connect study and Prolific study to test whether the 27 semantic antonyms were, in fact, antonymous. The Connect study involved administering the antonyms to a sample recruited from CloudResearch Connect (Hartman et al., 2023), and the Prolific study involved administering the antonyms to a sample recruited from Prolific. Given that both of these platforms are generally known for having higher-quality data than MTurk, we hypothesized that all of the semantic antonyms would manifest in negative correlations on the platforms (https://osf.io/wzf5u/?view_only=8c300b94bfe544948901bd4028370b35; https://osf.io/2g4ry/?view_only=9f4760a0e30c444b890408d256c5da84).

The goal of the open MTurk study was to test whether, despite manifesting in negative correlations on Connect and Prolific, the semantic antonyms would manifest in *positive* correlations on MTurk. We refer to this study as the *open* MTurk study because we recruited participants directly from MTurk without instituting any filters (see Stagnaro et al., 2024). Given prior research raising concerns about the quality of unfiltered data on MTurk (Douglas et al., 2023; Moss et al., 2021; Peer et al., 2022; Stagnaro et al., 2024; Webb & Tangney, 2022), as well as the prior study that found positive correlations for three semantic antonyms in this pool (Kay, 2024), we hypothesized that the 27 semantic antonyms would be positively correlated on the platform (https://osf.io/wzf5u/?view_only=8c300b94bfe544948901bd4028370b35). As part of this study, we also wanted to test whether any of the issues we observed could be addressed by implementing typical data-screening procedures. To do so, we further examined the associations between the semantic antonyms in the sample after excluding participants according to three common attention check measures. We did not have any hypotheses on this front.

The goal of the qualified MTurk study was to test whether any of the data quality issues we observed could be addressed by recruiting only so-called "high-productivity" and "high-reputation" MTurk participants. To test this, we administered the set of semantic antonyms to MTurk participants who had completed over 500 tasks on MTurk and achieved a task approval rate greater than 95% (Peer et al., 2014). We were doubtful that these qualifications would be sufficient to address the observed data quality issues, so we hypothesized that the antonyms would remain positively correlated (https://osf.io/8rxz7/?view_only=15e64202d55e4df183b44b3517f60083). As in the open MTurk study, we further tested whether the observed issues could be addressed by implementing common data-screening procedures. Again, we had no hypotheses on this front. The materials, data, and analytic code for all four studies can be found on OSF (https://osf.io/frwq4/?view_only=2e9d981a2df1411d8971a120ae75df06).

Method

Participants & procedures

Connect

The sample for the Connect study was collected on April 15, 2024. It was comprised of 100 participants who completed a Qualtrics survey posted to the on-demand data collection platform CloudResearch Connect. A power analysis indicated that 79 participants would be required to detect a correlation of .50 or larger (the smallest correlation deemed to be of practical interest) 95% of the time that such an effect existed in the population with an alpha level of .0017. The alpha level of .0017 was calculated by dividing the conventional alpha level of .05 by the total number of correlations we initially planned to test (Bonferroni, 1936).[2] We decided to collect 100 participants to allow for some misspecification in our power analysis. The participants were recruited from the US and paid $1.15 for their participation, a rate equivalent to $8.00 per hour. They ranged in age from 18 to 72 (*M* age = 36.71; *SD* age = 11.09) and mostly identified as women (53.00%), with 46.00% identifying as men, and 1.00% preferring not to answer the question. The participants were mostly white (63.00%), with a smaller number of participants identifying as Black (15.00%) and Asian (10.00%).

Prolific

The sample for the Prolific study was collected on July 28, 2024. It was comprised of 100 participants who completed the same Qualtrics survey administered to the Connect sample. Instead of recruiting participants from Connect, however, we recruited participants from Prolific. We based our desired sample size on the same power analysis used for the Connect study. The participants were

[2] We initially planned to test the associations between 29 pairs of items but, in the end, only tested the associations between 27. Upon reviewing the results from our Connect sample, we discovered that one pair of items from our initial set of 28 semantic antonyms was actually a pair of semantic *synonyms* (e.g., "I seldom feel blue" and "I am generally a happy person"; $r = .48$, $p <. 001$). Because this pair of items was not able to serve its intended purpose, we dropped it from further analysis. We had also initially included a pair of gibberish items (i.e., "I am ffhjhl" and "I am sqnmmp") in the survey under the assumption that the items would not be associated with each other in a sample of careful respondents. This was not the case. The items were highly associated in the Connect ($r = .99$, $p < .001$) and Prolific ($r = .84$, $p < .001$) samples. In hindsight, this outcome was obvious. How a careful participant responds to one inscrutable item is likely going to be similar to how they respond to a second inscrutable item. Again, since this pair of items was not able to serve its intended purpose, we dropped it from further analysis.

recruited from the US and paid $1.15 for their participation. They ranged in age from 18 to 73 (*M* age = 33.44; *SD* age = 11.51) and mostly identified as women (55.00%), with 38.00% identifying as men, and 7.00% providing some other response. The participants were mostly white (58.00%), with a smaller number of participants identifying as Black (20.00%) and Asian (8.00%).

Open MTurk

The sample for the open MTurk study was collected between April 15, 2024, and April 17, 2024. It was comprised of 400 MTurk participants who completed the same Qualtrics survey administered in the other two studies. A power analysis indicated that 242 participants would be required to detect a correlation of .30 or larger (the smallest correlation deemed to be of practical interest) 95% of the time that such an effect existed in the population, with an alpha level of .0017. We decided to collect 400 participants to allow for some misspecification in our power analysis. The participants were, again, recruited from the US and paid $1.15 for their participation. They ranged in age from 18 to 69 (*M* age = 32.12; *SD* age = 6.98) and, in this case, mostly identified as men (77.25%), with 22.50% identifying as women and 0.25% identifying as gender fluid. The participants were mostly white (84.00%), with a smaller number of participants identifying as Asian (4.50%) and Hispanic or Latinx (3.50%).

Qualified MTurk

The sample for the qualified MTurk study was collected between November 17, 2024, and November 18, 2024. It was comprised of 600 MTurk participants who had completed over 500 tasks on the platform and had a task approval rate greater than 95%. They completed the same Qualtrics survey administered in the other three studies. Given we expected the correlations to be somewhat smaller in this qualified sample than in the open sample, we used a smaller correlation for our power analysis (i.e., $r = .20$). The power analysis indicated that 560 participants would be required to detect a correlation of this size or larger 95% of the time that such an effect existed in the population using our conservative alpha level of .0017. We decided to collect 600 participants to allow for some misspecification in our power analysis. As in the other studies, the participants were located in the US and paid $1.15 for their participation. They ranged in age from 22 to 72 (*M* age = 33.26; *SD* age = 7.80) and mostly identified as men (80.67%), with 19.00% identifying as women, 0.17% identifying as gender fluid, and 0.17% identifying as nonbinary. The participants were mostly white (77.67%), with a smaller number of participants identifying as Asian (7.00%), Black (4.50%), and Hispanic or Latinx (3.50%).

Materials

Participants in the four studies responded to 27 pairs of semantic antonyms.[3,4] Fourteen of the semantic antonyms were created by taking items from Donnellan and colleagues' (2006) highly abridged *Big-Five Factor Markers* (see also Goldberg, 1992)—often called the "Mini-IPIP"—and writing items that were antithetical to the original items. For example, the item "I like to be the center of attention" was created as a semantic antonym for the item "I keep in the background." Two of the original items (i.e., "I don't talk a lot"; "I am not really interested in others") were rewritten to avoid negatives (i.e., "I talk a lot"; "I am interested in others"). We did this to avoid a situation where a positive correlation could appear between two antonyms because participants accidentally missed the negating term. An additional five semantic antonyms were included from the *Semantic Antonyms Set* (see Kay, 2024). One of the items from the set was, again, reworded to remove a negative, and three of the other items were rewritten for clarity (e.g., "I am always suffering from an illness" became "It feels like I'm almost always suffering from a cold"). Finally, eight novel antonym pairs were created for the present study (e.g., "I'm a loner" and "I have many friends").

To investigate whether any positive associations observed between the semantic antonyms in the two MTurk samples could be addressed by simply screening one's data, we coded participants using three common indices of careless responding. First, we coded whether the participants (1) responded to the survey faster than 2 s per item (Huang et al., 2012); (2) responded the same way to over half of the survey items in a row (Johnson, 2005); and (3) responded incorrectly to two or more of four instructed response items embedded in the survey (e.g., "Choose strongly disagree for this item"; see Curran, 2016). Importantly, we did not choose these indices because they represent a comprehensive accounting of the various indices of careless responding, nor

[3] Descriptive statistics for the 54 items can be found in the Supplementary Material.

[4] Since we were primarily interested in platform-level effects—whether the semantic antonyms were positively correlated on MTurk—rather than individual-level effects—whether the semantic antonyms were positively correlated for a given participant—we report between-person correlations rather than within-person correlations here. Nevertheless, for each of the four samples, we also calculated the percentage of participants who would be flagged for careless responding according to various within-person correlation thresholds (see the Supplementary Material). Consistent with the results reported here, nearly half of those in the open MTurk (47.56%) and qualified MTurk (42.68%) samples exhibited positive correlations between the semantic antonyms. In contrast, almost no one in the Connect (1.01%) and Prolific (0.00%) samples exhibited positive correlations between the semantic antonyms.

because they comprise the best indices for detecting careless responding. We chose these indices because they align with what we expect the average researcher would reasonably use (but see Stosic et al., 2024). For those interested, we have included the code to calculate additional indices of careless responding, including intra-individual response variability (Thalmayer & Saucier, 2014; see also Dunn et al., 2018), horizontal cursor travel (Pokropek et al., 2023), and personal-total correlations (Donlon & Fischer, 1968; see also Curran, 2016) on OSF (https://osf.io/frwq4/?view_only=2e9d981a2df1411d8971a120ae75df06).[5]

Results

Connect

Most (77.78%) of the 27 semantic antonyms were negatively correlated at our conservative alpha level of .0017 in the Connect sample (Table 1; Fig. 1; see the Supplementary Material for scatterplots). In fact, there were only six exceptions: (1) "I never get sick" was not significantly negatively associated with "It feels like I'm almost always suffering from a cold" (r=–.25, p= .011); (2) "I am interested in abstract ideas" was not significantly negatively associated with "I only concern myself with concrete ideas" (r=–.28, p= .004); (3) "I like order" was not significantly negatively associated with "I crave chaos" (r=–.24, p= .017); (4) "I have many pets" was not significantly negatively associated with "I have few pets" (r=–.15, p= .130); (5) "I am interested in others" was not significantly negatively associated with "Most people bore me" (r=–.28, p= .004); and (6) "I like to play it safe" was not significantly negatively associated with "Danger excites me" (r=–.31, p= .002). Some of these exceptions may simply be false negatives (i.e., type II errors), but it is also possible that some of the content may not be as antonymous as we initially believed. For example, one can easily imagine a person who likes to play it safe but, nevertheless, finds danger exciting. Whatever the case may be, the present results indicate that the majority of the semantic antonyms tested here capture opposing content.

Prolific

Similar to the Connect study, most (85.19%) of the semantic antonyms were negatively correlated at our conservative alpha level of .0017 in the Prolific sample (Table 1; Fig. 1; see the Supplementary Material for scatterplots). In this case, there were only four exceptions: (1) "I am narcissistic" was not significantly negatively associated with "I am a selfless person" (r=–.13, p= .205); (2) "I am interested in abstract ideas" was not significantly negatively associated with "I only concern myself with concrete ideas" (r=–.21, p= .033); (3) "I have many pets" was not significantly negatively associated with "I have few pets" (r=–.05, p= .612); and (4) "I make a mess of things" was not significantly negatively associated with "I improve most things I touch" (r=–.28, p= .005). As in the Connect study, some of these associations may simply be false negatives, but, again, it is possible that some of the content may not be as antonymous as we initially believed. Nevertheless, the results again indicate that the majority of the semantic antonyms are empirically antonymous.

Open MTurk

In contrast to the results of the Connect and Prolific studies, nearly all (96.30%) of the semantic antonym pairs were significantly *positively* correlated in the open MTurk sample (Table 1; Fig. 1; see the Supplementary Material for scatterplots). The one semantic antonym pair that was not significantly positively correlated was "I consider myself a Democrat" and "I consider myself a Republican" (r= .09, p= .078), although these items were still not significantly negatively correlated as they were in the Connect (r=–.64, p < .001) and Prolific (r=–.64, p < .001) samples. We suspect that participants may have been more attentive (and, therefore, more likely to respond correctly) to these two items because they were the only items that included capital letters. We also suspect that participants may have believed that these two items would qualify them for future studies and, consequently, took additional pains to respond accurately. Whatever the case may be, the results indicate that a large number of participants on MTurk respond in similar ways to items assessing contradictory content,

[5] The data for the four studies also includes responses to a question asking participants whether they are human and a question asking participants whether they are a large language model. The responses to these questions are not directly relevant to the present project, but it is notable that, in response to the question of whether they are a large language model, 37.00% of the participants in the open MTurk sample said "yes"; 1.75% said "very large"; 3.00% provided some definition of a large language model (e.g., "A type of artificial intelligence (AI) program that can recognize and generate text, among other tasks"); 27.25% said "English" (perhaps because they interpreted the question as asking them what language they spoke); and 0.25% said "Spanish" (see prior parenthetical). Even among those in the qualified MTurk sample, 22.50% responded with some variant of "yes"; 1.50% simply said "llm"; 0.33% provided some definition of a large language model; and 34.17% said "English." In contrast, 100.00% of the participants in the Connect sample and 92.00% of the participants in the Prolific sample said some variant of "no." Of the 8.00% of respondents in the Prolific sample who did not say "no", 3.00% responded "yes" and 5.00% expressed confusion about what a large language model is. Correlations between the semantic antonyms in the open MTurk and qualified MTurk samples after excluding participants who did not say some variant of "no" can be found in the Supplementary Material.

Table 1 Correlations between 27 semantic antonyms in a (1) Connect sample ($N_1 = 100$), (2) Prolific sample ($N_2 = 100$), (3) open MTurk sample ($N_3 = 400$), (4) open MTurk sample with participants screened out using common attention check measures ($N_{3\text{-Screened}} = 211$), (5) qualified MTurk sample ($N_4 = 600$), and (6) qualified MTurk sample with participants screened out using common attention check measures ($N_{4\text{-Screened}} = 232$)

				Open MTurk		Qualified MTurk	
Item 1	Item 2	Connect	Prolific	Not Screened	Screened	Not Screened	Screened
I never get sick	It feels like I'm almost always suffering from a cold	–.25*	–.36**	.62**	.62**	.53**	.48**
I get upset easily	It takes a lot to upset me	–.64**	–.69**	.58**	.51**	.63**	.63**
I talk to a lot of different people at parties	I only really feel comfortable talking with people I know	–.54**	–.40**	.54**	.48**	.45**	.20*
I am narcissistic	I am a selfless person	–.34**	–.13	.51**	.45**	.42**	.24**
I sleep soundly	I find it hard to get a good night's rest	–.83**	–.66**	.49**	.35**	.42**	.38**
I am interested in abstract ideas	I only concern myself with concrete ideas	–.28*	–.21*	.49**	.29**	.48**	.22**
I talk a lot	I rarely talk	–.56**	–.48**	.46**	.29**	.27**	.02
I keep in the background	I like to be the center of attention	–.56**	–.39**	.46**	.29**	.37**	.16*
I get chores done right away	I put off my chores until the last minute	–.66**	–.67**	.46**	.32**	.45**	.36**
I have frequent mood swings	My mood tends to be stable	–.66**	–.74**	.45**	.27**	.38**	.08
I like order	I crave chaos	–.24*	–.44**	.45**	.25**	.38**	.25**
I am good at saving my money	I go through money quickly	–.67**	–.69**	.44**	.25**	.38**	.19*
I have many pets	I have few pets	–.15	–.05	.43**	.25**	.24**	.05
I make a mess of things	I improve most things I touch	–.42**	–.28*	.42**	.32**	.35**	.17*
I feel others' emotions	I often find it difficult to tell how other people are feeling	–.46**	–.52**	.40**	.18*	.36**	.14*
I stay up-to-date with current events	I rarely follow the news	–.73**	–.61**	.40**	.21*	.30**	.02
I have a vivid imagination	I find it difficult to imagine things	–.48**	–.50**	.40**	.23**	.38**	.15*
I sympathize with others' feelings	I am indifferent to the feelings of others	–.65**	–.35**	.40**	.28**	.39**	.14*
I am relaxed most of the time	I am constantly on edge	–.59**	–.54**	.39**	.25**	.31**	.11
I adapt easily to new situations	I am uncomfortable with things that are new to me	–.50**	–.54**	.37**	.24**	.33**	.12
I would never take something that didn't belong to me	I would steal things if given the chance	–.62**	–.48**	.37**	.14*	.31**	.17*
I am interested in others	Most people bore me	–.28*	–.37**	.36**	.18*	.33**	.05
I find it easy to stay focused	I am easily distracted	–.65**	–.56**	.34**	.15*	.24**	.02
I like to play it safe	Danger excites me	–.31*	–.37**	.32**	.14*	.32**	.10
I am an extrovert	I am an introvert	–.85**	–.75**	.32**	.09	.19**	–.12
I have many friends	I'm a loner	–.64**	–.50**	.24**	–.01	.24**	–.13*
I consider myself a Democrat	I consider myself a Republican	–.64**	–.64**	.09	–.11	.10*	–.23**

* $p < .05$, ** $p < .0017$. In the screened samples, participants were excluded if they responded faster than two seconds per item, provided the same response to over half of the items on the survey in a row, and/or failed two or more of the four instructed response items embedded in the survey. The MTurk participants in the qualified sample had all completed over 500 tasks on MTurk and had a task approval rate greater than 95%

suggesting that there is, in fact, an issue with data quality on the platform.

Our results also indicated that the basic screening of one's data is unlikely to be sufficient to extract usable data from an open MTurk sample. Even after we excluded 47.25% of the sample for failing one or more of the three common attention check measures, 66.67% of the semantic antonyms remained significantly positively correlated (Table 1; Fig. 1; see the Supplementary Material for scatterplots).

Qualified MTurk

As in the open MTurk study, nearly all (96.30%) of the semantic antonyms were significantly positively correlated

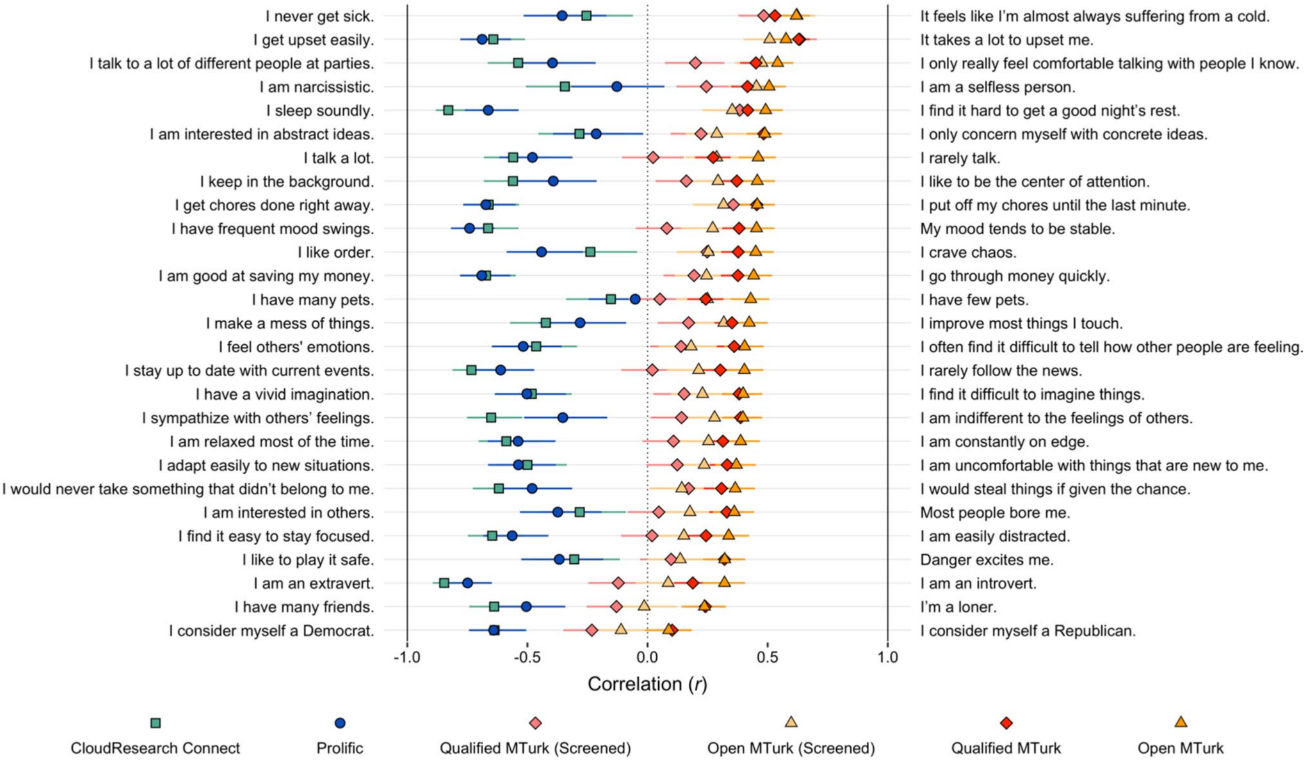

Fig. 1 Correlations between 27 semantic antonyms in a (1) Connect sample ($N_1 = 100$), (2) Prolific sample ($N_2 = 100$), (3) open MTurk sample ($N_3 = 400$), (4) open MTurk sample with participants screened out using common attention check measures ($N_{3\text{-Screened}} = 211$), (5) qualified MTurk sample ($N_4 = 600$), and (6) qualified MTurk sample with participants screened out using common attention check measures ($N_{4\text{-Screened}} = 232$). Error bars represent 95% confidence intervals. In the screened samples, participants were excluded if they responded faster than two seconds per item, provided the same response to over half of the items on the survey in a row, and/or failed two or more of the four instructed response items embedded in the survey. The MTurk participants in the qualified sample had all completed over 500 tasks on MTurk and had a task approval rate greater than 95%

at our conservative alpha level of .0017 in the qualified MTurk sample (Table 1; Fig. 1; see the Supplementary Material for scatterplots). The one semantic antonym pair that did not show a significant positive association was, again, "I consider myself a Democrat" and "I consider myself a Republican" ($r = .10$, $p = .012$). Taken together, these results indicate that even so-called "high productivity" and "high reputation" participants on MTurk are not providing usable data.

Screening the data using common attention check measures also did not appear to be sufficient to extract usable data from the qualified MTurk sample. After excluding 61.33% of the sample using our three common attention check measures, 25.93% of the semantic antonyms remained significantly positively correlated (Table 1; Fig. 1; see the Supplementary Material for scatterplots), and only one of the semantic antonyms became significantly negatively correlated.[6] This one exception was "I consider myself a Democrat" and "I consider myself a Republican" ($r = –.23$, $p < .001$).

Discussion

Several prior studies have employed advanced methodological and statistical techniques to highlight issues with the quality of data that can be collected on MTurk (Douglas et al., 2023; Moss et al., 2021; Peer et al., 2022; Stagnaro

[6] Interestingly, the proportion of participants excluded using the three common attention check measures was greater in the qualified MTurk sample (61.33%) than in the open MTurk sample (47.25%). This seems to suggest that many qualified MTurk participants are regularly engaging in careless responding but not having their submissions rejected for doing so.

et al., 2024; Webb & Tangney, 2022). The purpose of the present set of studies was to provide an accessible demonstration of this issue using a face-valid indicator of data quality: Do items that assess clearly contradictory content show positive correlations on the platform?

The results were clear. Despite most of the items manifesting in negative correlations on Connect (77.78%) and Prolific (85.19%), 96.30% of the semantic antonyms were *positively* correlated in an open sample of MTurk participants and a qualified sample of "high-productivity", "high-reputation" MTurk participants. This means that a large number of participants on MTurk responded in a similar way to items like "I never get sick" and "It feels like I'm almost always suffering from a cold"; "I get upset easily" and "It takes a lot to upset me"; and "I talk to a lot of different people at parties" and "I only really feel comfortable talking with people I know." This pattern remained even when screening the data using common attention check measures. After screening out nearly half of the participants from the open MTurk sample, 66.67% of the correlations remained significantly positively correlated, and, after screening out nearly two-thirds of the qualified MTurk sample, 25.93% of the correlations remained significantly positively correlated. Taken together, the present results indicate that the average dataset collected on MTurk simply cannot be trusted.

The conclusion that data collected on MTurk cannot be trusted is, of course, troubling. There are many benefits to using online data-collection platforms like MTurk (see Reips, 2000), including being able to collect data more efficiently and having access to samples that are more diverse than those typically used in psychological research (Arnett, 2008; Thalmayer et al., 2021). However, if we found positive correlations between diametrically opposed items in the present MTurk samples, there is no reason to believe we wouldn't also find a positive correlation between almost any other pair of items. Studies that use data collected on MTurk can, therefore, find associations between constructs that are, quite literally, counter to reality. The consequences of this should not be understated. Not only can such spurious results cause researchers to waste time, effort, and funds pursuing lines of research that are demonstrably false, but they can also threaten a researcher's ability to generate accurate knowledge about the world, undermining the very goal of scientific research.

The next question, then, is what can be done? One option is to be even more circumspect about the participants we choose to retain in our MTurk samples. However, the relatively liberal inclusion criteria used here removed approximately half of our open MTurk sample and approximately two-thirds of our qualified MTurk sample, and prior studies using more stringent inclusion criteria have resulted in the exclusion of nearly all of their participants (Webb & Tangney, 2022; but see Cuskley & Sulik, 2024). Under these conditions, a researcher could easily deplete all of their research funds before achieving a desired sample size. Moreover, even if researchers withheld funds for all of the participants they excluded, the time required to repeatedly collect, screen, and recollect participants would be significant. We, therefore, do not find this first option to be a tenable solution.

Instead of changing one's inclusion criteria, a second option is to try to implement interventions that would decrease the incidence of careless responding on MTurk to begin with. However, many of these interventions do not appear to be particularly effective (e.g., Brühlmann et al., 2023; Marshall, 2019). Until effective interventions are developed, we can't, in good conscience, recommend this as a reasonable option.

A third (and potentially less popular) option is to call for a moratorium on running MTurk studies. Although this wasn't true in the early days of MTurk, there are now a number of on-demand data collection platforms (e.g., CloudResearch Connect; Prolific) that can be used to collect data instead of MTurk, and many of these platforms do not appear to suffer from the same data quality issues as MTurk (Douglas et al., 2023; Stagnaro et al., 2024). The common objection to platforms like Connect and Prolific is that they cost more than MTurk. This is true. Connect requires a minimum payment of $6.50 per hour, and Prolific requires a minimum payment of $8.00 per hour. In contrast, MTurk requires a minimum payment of $0.01 *per assignment*. We have two responses to this objection.

First, when evaluating the cost of a project, one needs to consider the cost *per high-quality respondent,* which can vary considerably across platforms. While Connect and Prolific regularly screen their users for evidence of careless responding, MTurk seems to either not screen its users or screens its users only minimally. As a result, a participant recruited from MTurk is, on average, more likely to be a careless respondent than a participant recruited from Connect or Prolific. Researchers using MTurk must, therefore, recruit a larger number of participants to achieve samples with the same number of high-quality respondents as researchers using either of the other two platforms, driving up the average cost for each participant that can actually be used in one's analyses. As a case in point, Douglas and colleagues (2023) found that they needed to spend $4.36 for each high-quality respondent they collected on MTurk while they only needed to spend $1.90 for each high-quality respondent they collected on Prolific. Of course, these numbers were calculated assuming the same base pay on both services, which is often not the case in practice. Oftentimes, researchers award participants far less on MTurk. Nonetheless, considering the cost per high-quality respondent reveals that the overall administration costs on these platforms are closer than they initially appear.

Our second response is simply that researchers should consider whether paying participants more is, in fact, a bad thing. We understand that researchers will disagree on what is an ethical amount to pay participants, but it is troubling how little some MTurk participants are paid. One estimate indicates that 61% of the assignments on MTurk pay less than $0.10, with 52% of MTurk participants reporting that they make less than $5.00 per hour (Pew Research Center, 2016; but see Moss et al., 2023). Given these numbers, it is perhaps unsurprising that 13% of MTurk participants feel that they do not receive fair pay for their work (Fowler et al., 2023) and 59% of MTurk participants feel that they are being exploited to some degree (Busarovs, 2013), with this number increasing to 64% if one defines "exploitation" before asking the question. We encourage researchers to pay participants, at the very least, minimum wage (see Gleibs, 2017). If researchers choose to follow this advice, the one benefit of using MTurk over other on-demand data collection platforms (i.e., its cost) all but disappears.

Limitations and future directions

The present study was not without its limitations. First, we assessed a relatively small portion of the over 500,000 registered (but potentially not active; Fort et al., 2011) workers on MTurk (Amazon Web Services, 2024). Although we have no reason to believe that the sample we collected would be different from any other sample collected on MTurk, we nonetheless encourage researchers to conduct replications of our results using larger numbers of MTurk workers. Second, all of our samples were collected on either a single date or within a narrow range of dates in 2024. Again, we have no reason to believe that samples collected on other dates would have yielded different results than those observed here, except perhaps samples collected many years ago, before MTurk started to suffer from the data quality issues that it does today (see Peer et al., 2014). Still, we encourage researchers to conduct replications of our results using samples drawn from a broader range of dates, both to evaluate the robustness of our findings and to track how the quality of data on MTurk is changing over time. Finally, our test of screening procedures used a relatively permissive set of inclusion criteria. This was, of course, intentional: other studies have shown what happens when one uses a more stringent set of inclusion criteria (Webb & Tangney, 2022; but see Cuskley & Sulik, 2024), and we wanted to examine the quality of data on MTurk after screening participants using procedures that the average researcher would be likely to use. Nevertheless, we encourage future work focused on evaluating the efficacy of different data screening procedures on the platform.

Beyond addressing the limitations above, we also recommend the development of additional interventions to combat careless responding on MTurk. As we noted above, existing interventions aimed at reducing careless responding on MTurk have not been especially effective (e.g., Brühlmann et al., 2024; Marshall, 2019), but this does not mean an intervention could not be identified that could decrease careless responding on the platform. The present findings could even aid in such efforts. The methodology employed here provides a coarse but face-valid index of the quality of data on MTurk, which could be used to test the effectiveness of novel interventions. Does an intervention result in negative correlations among semantic antonyms on MTurk? If the answer is "yes", it suggests that the intervention is effective. If the answer is "no", it suggests the intervention is ineffective.

Conclusion

The present project examined the associations between 27 semantic antonyms (e.g., "I like order" and "I crave chaos") on MTurk. A large number of the antonyms were positively correlated on the platform, even after the data was subjected to typical data screening procedures and even when only "high-productivity" and "high-reputation" participants were sampled. These results are as concerning as they are unequivocal: the average dataset collected on MTurk simply cannot be trusted.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-025-02852-7.

Acknowledgements I would like to thank Gabby Gavala for her help during the initial stages of this project, as well as Union College for funding all of the studies reported here.

Author contributions Cameron Kay: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing—original draft, Writing—review & editing.

Funding Union College funded all of the studies reported here.

Data availability The data and materials for the four studies can be found on OSF (https://osf.io/frwq4/?view_only=2e9d981a2df1411d8971a120ae75df06).

Code availability The analytic code for the four studies can be found on OSF (https://osf.io/frwq4/?view_only=2e9d981a2df1411d8971a120ae75df06).

Declarations

Conflicts of interest/Competing interests None.

Ethics approval The study reported here was determined to be exempt by the Human Subjects Review Committee at Union College (E24016;

E24043). It was conducted in accordance with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent to participate Participants provided informed consent at the beginning of all of the surveys reported here.

Consent for publication Not applicable.

Open Practices Statement The materials, data, and analytic code for the four studies can be found on OSF (https://osf.io/frwq4/?view_only=2e9d981a2df1411d8971a120ae75df06). All four studies were preregistered. The preregistration for the Connect study can be found on OSF (https://osf.io/wzf5u/?view_only=8c300b94bfe544948901bd4028370b35). The preregistration for the Prolific study can be found on OSF (https://osf.io/2g4ry/?view_only=9f4760a0e30c444b890408d256c5da84). The preregistration for the open MTurk study can be found on OSF (https://osf.io/wzf5u/?view_only=8c300b94bfe544948901bd4028370b35). The preregistration for the qualified MTurk study can be found on OSF (https://osf.io/8rxz7/?view_only=15e64202d55e4df183b44b3517f60083).

References

Amazon Web Services. (2024). *Amazon Mechanical Turk: Requester UI Guide*.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63*(7), 602–614.

Bakan, P. (1960). Response-tendencies in attempts to generate random binary series. *American Journal of Psychology, 73*(1), 127–131.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351–368.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8*, 3–62.

Brühlmann, F., Memeti, Z., Aeschbach, L. F., Perrig, S. A. C., & Opwis, K. (2024). The effectiveness of warning statements in reducing careless responding in crowdsourced online surveys. *Behavior Research Methods*, *2018*.

Busarovs, A. (2013). Ethical aspects of crowdsourcing, or is it a modern form of exploitation. *International Journal of Economics and Business Administration, 1*(1), 3–14.

Butcher, J. N., Cox, A. C., Weed, N. C., & Butcher, J. N. (2009). The MMPI-2: History, interpretation, and clinical issues. In *Oxford Handbook of Personality Assessment* (pp. 250–276).

Cornell, D., Klein, J., Konold, T., & Huang, F. (2012). Effects of validity screening items on adolescent survey data. *Psychological Assessment, 24*(1), 21–35.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*(4), 596–612.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal Of Experimental Social Psychology, 66*, 4–19.

Cuskley, C., & Sulik, J. (2024). The burden for high-quality online data collection lies with researchers, not recruitment platforms. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916241242734

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology, 67*(2), 309–338.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement, 28*, 105–113.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment, 18*(2), 192–203.

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), Article e0279720.

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.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*(2), 301–318.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics, 37*(2), 413–420.

Fowler, C., Jiao, J., & Pitts, M. (2023). Frustration and ennui among Amazon MTurk workers. *Behavior Research Methods, 55*(6), 3009–3025. https://doi.org/10.3758/s13428-022-01955-9

Gleibs, I. H. (2017). Are all "research fields" equal? Rethinking practice for the use of data from crowdsourcing market places. *Behavior Research Methods, 49*(4), 1333–1342. https://doi.org/10.3758/s13428-016-0789-y

Goldberg, L. R. (1992). Development of markers for the big-five factor structure. *Psychological Assessment, 4*(1), 26–42.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48*(1), 82–98.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research, 44*(1), 196–210.

Hartman, R., Moss, A. J., Jaffe, S. N., Rosenzweig, C., Litman, L., & Robinson, J. (2023). Introducing Connect by CloudResearch: Advancing online participant recruitment in the digital age. https://doi.org/10.31234/osf.io/ksgyr

Holtzman, N. S., & Donnellan, M. B. (2017). A simulator of the degree to which random responding leads to biases in the correlations between two individual differences. *Personality and Individual Differences, 114*, 187–192.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal Of Applied Psychology, 100*(3), 828–845.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal Of Applied Psychology, 75*(5), 581–595.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*, 103–129.

Kay, C. S. (2024). Validating the IDRIS and IDRIA: Two infrequency/frequency scales for detecting careless and insufficient effort survey responders. *Behavior Research Methods*, 1–24. https://doi.org/10.31234/osf.io/us7bm

Marshall, A. D. (2019). *Applying the theory of planned behavior to participation*. [Doctoral dissertation]. Colorado State University.

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. *Preprint*. https://doi.org/10.31234/osf.io/wr8ds

Moss, A. J., Rosenzweig, C., Robinson, J., Jaffe, S. N., & Litman, L. (2023). Is it ethical to use Mechanical Turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages. *Behavior Research Methods, 55*(8), 4048–4067.

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.

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

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*, 1643–1662.

Pew Research Center. (2016). *Research in the crowdsourcing age, a case study*. https://www.pewresearch.org/internet/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

Pokropek, A., Żółtak, T., & Muszyński, M. (2023). Mouse chase: Detecting careless and unmotivated responders using cursor movements in web-based surveys. *European Journal of Psychological Assessment, 39*(4), 299–306.

Porter, C. O. L. H., Outlaw, R., Gale, J. P., & Cho, T. S. (2019). The use of online panel data in management research: A review and recommendations. *Journal of Management, 45*(1), 319–344.

Reips, U.-D. (2000). The web experiment: Advantages, disadvantages, and solutions. *Psychology Experiments on the Internet, 1995*, 89–117.

Stagnaro, M. N., Druckman, J. N., Berinsky, A. J., Arechar, A. A., Willer, R., & Rand, D. G. (2024). *Representativeness versus attentiveness: A comparison across nine online survey samples*. 1–43.

Stosic, M. D., Murphy, B. A., Duong, F., Fultz, A. A., Harvey, S. E., & Bernieri, F. (2024). Careless responding : Why many findings are spurious or spuriously inflated. *Advances in Methods and Practices in Psychological Science, 7*(1), 1–19.

Thalmayer, A. G., & Saucier, G. (2014). The questionnaire Big Six in 26 nations: Developing cross-culturally applicable Big Six, Big Five and Big Two inventories. *European Journal of Personality, 28*, 482–496. https://doi.org/10.1002/per.1969

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist, 76*(1), 116–129. https://doi.org/10.1037/amp0000622

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

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111*(4), 493–504.

Zorowitz, S., Solis, J., Niv, Y., & Bennett, D. (2023). Inattentive responding can induce spurious associations between task behavior and symptom measures. *Nature Human Behaviour, 7*, 1667–1681.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.