2022

Collecting Samples From Online Services: How to Use Screeners to Improve Data Quality

Aaron D. Arndt
Old Dominion University, aarndt@odu.edu

John B. Ford
Old Dominion University, jbford@odu.edu

Barry J. Babin

Vinh Luong
Old Dominion University, vluon001@odu.edu

Original Publication Citation

Arndt, A. D., Ford, J. B., Babin, B. J., & Luong, V. (2022). Collecting samples from online services: How to use screeners to improve data quality. *International Journal of Research in Marketing, 39*(1), 177-133. https://doi.org/10.1016/j.ijresmar.2021.05.001

International Journal of Research in Marketing 39 (2022) 117–133

Contents lists available at ScienceDirect

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full length article

Collecting samples from online services: How to use screeners to improve data quality

Aaron D. Arndt[a,*], John B. Ford[b], Barry J. Babin[c], Vinh Luong[d]

[a] *Old Dominion University, Strome College of Business, Department of Marketing, 2055 Constant Hall, Norfolk, VA 23529, United States*
[b] *Old Dominion University, Strome College of Business, Department of Marketing, 2117 Constant Hall, Norfolk, VA 23529, United States*
[c] *University of Mississippi, Ole Miss Business School, Department of Marketing, 235 Holman Hall University, MS 38677, United States*
[d] *Old Dominion University, Strome College of Business, Department of Marketing, Constant Hall, Norfolk, VA 23529, United States*

ARTICLE INFO

Article history:
First received on March 23, 2020 and was under review for 8 months
Available online 17 May 2021

Area Editor: Donald R. Lehmann

Keywords:
Online samples
Sampling
MTurk
Screening
Data quality
Integrity

ABSTRACT

Increasingly, marketing and consumer researchers rely on online data collection services. While actively-managed data collection services directly assist with the sampling process, minimally-managed data collection services, such as Amazon's Mechanical Turk (MTurk), leave researchers solely responsible for recruiting, screening, cleaning, and evaluating responses. The research reported here proposes a 2 × 2 framework based on sampling goal and methodology for screening and evaluating the quality of online samples. By sampling goals, screeners can be categorized as *selection*, which involves matching the sample with the targeted population; or as *accuracy*, which involves ensuring that participants are appropriately attentive. By methodology, screeners can be categorized as *direct*, which screens individual responses; and as *statistical*, which provides quantitative signals of low quality. Multiple screeners for each of the four categories are compared across three MTurk samples, two actively-managed data collection samples (Qualtrics and Dynata), and a student sample. The results suggest the need for screening in every online sample, particularly for the MTurk samples, with the fewest supplier-provided filters. Recommendations are provided for researchers and journal reviewers that provide greater transparency with respect to sample practices.

1. Introduction

An increasing number of marketing researchers are buying data through online data collection services (Goodman & Paolacci, 2017). An examination of four prominent journals in Marketing (*Journal of Marketing*, *Journal of Marketing Research*, *Journal of Consumer Research* and *Journal of the Academy of Marketing Science*) from 2014 to 2018 indicates that 360 out of 1101 articles (32.7%) purchased responses from Amazon's online crowdsourcing website, Mechanical Turk (aka, MTurk). Yet, only about 16% of these studies report procedures used to achieve data quality. To date, marketing research focuses on comparing the reliability and generalizability of data obtained through online crowdsourcing to other data sources (Kees, Berry, Burton, & Sheehan, 2017) and assessing the veracity of responses from MTurk (Wessling, Huber, & Netzer, 2017). Little attention is given to the effectiveness of screening procedures used to assess and control overall online survey data quality. For example, Wessling et al. (2017) explain that many MTurk respondents (aka, MTurkers) provide low-quality

[*] Corresponding author.
E-mail addresses: aarndt@odu.edu (A.D. Arndt), jbford@odu.edu (J.B. Ford), bbabin@bus.olemiss.edu (B.J. Babin), vluon001@odu.edu (V. Luong).

https://doi.org/10.1016/j.ijresmar.2021.05.001
0167-8116/© 2021 The Authors. Published by Elsevier B.V.

data, but they provide only limited practical guidance for managing data quality. Researchers are faced with a myriad of choices for assessing and controlling data quality. An empirical evaluation of data screening procedures would help researchers to understand the impact of their choices.

Screening is particularly important for crowdsourced data collection services like MTurk because they are minimally-managed. Traditional sampling services, such as Nielsen and Dynata, actively-manage panel membership. They recruit and verify the identity of panel members and then use statistical techniques to draw samples from their membership panels. By contrast, minimally-managed data collection services recruit internet workers to "work" by responding to online posts inviting survey participants. The effort made to verify participant identity is minimal. An MTurk work request is analogous to putting a billboard up asking people to send their opinions on certain matters for pay. Respondents self-select (single opt-in) into studies with the resulting sample highly unlikely to be representative of a relevant target population. Because research materials must be customized to a limited set of languages and cultures, the sample representativeness of a target population is important for all human-subjects research, even for theory that is universally generalizable. Nonrepresentative samples become sources of bias when members interpret experimental manipulations and measurement scales differently relative to the intended population. Thus, all researchers should have the capability to estimate representativeness so they can make informed decisions.

The goal of scientific sampling in human subjects' research is to collect a sample that matches the target population characteristics as much as possible from respondents who are appropriately attentive and honest (Berinsky, Huber, & Lenz, 2012). While actively-managed data collection services help researchers ensure response quality, minimally-managed data collection services provide only built-in options for prescreening data and are not otherwise involved in the data collection process. Researchers using minimally-managed services have a dizzying number of options for screening samples. MTurk provides options allowing a requester to select only *accounts* with a minimum number of jobs (Human Intelligence Tasks or HITs) completed, minimum approval rate (the percentage of HITs that requesters felt were completed adequately), reported location, or by choosing to make the HIT request available only to MTurk Masters (MTurker accounts completing an extremely large number of HITs with a high approval rating can receive the Master designation).[1] Additionally, researchers can include an almost infinite number of screening criteria in the research materials to select respondents and ensure that they are paying attention, such as requiring certain respondent demographics (e.g., female) or correct answers to attention checks (e.g., pick choice "2″).

Currently, the marketing literature provides very little empirically-based guidance as to how to choose, build, assess, or report screening criteria for online samples. DeSimone, Harms, and DeSimone (2015) categorize screening techniques by methodology, but they do not provide actionable recommendations for when or how to use screeners to improve data quality. Thus, marketing researchers lack clear guidelines for screening data. The purpose of this research is to: (1) establish a typology for classifying screening criteria and (2) evaluate the number of responses screened by criteria for each data source. In other words, what are effective screeners and how do they work? Such insight should help researchers who opt to use online samples to better evaluate response validity and, also, to help reviewers evaluate sample quality more effectively. Ultimately, data generalizability is necessary for external validity and, subsequently, the development of meaningful managerial implications.

2. Background on online crowdsourcing samples

Traditional survey techniques, such as telephone, mall-intercept, and postal-mail interviews, have been largely replaced in academic marketing research by online approaches, often involving fee-based data access obtained through intermediaries. Yet, online data sources are not homogeneous. Marketing research services like Dynata and Kantar, and online survey technology firms like Qualtrics, offer access to actively-managed pools of respondents, in which respondents should be vetted and prescreened before signing up to receive survey invitations. Because respondents cannot change their identity without going through a long process, and they cannot know what surveys they are "missing," panel members have far fewer opportunities or incentives to mislead researchers. Other services, such as MTurk, provide access to minimally-managed pools of respondents and delegate respondent screening to researchers. Thus, the same person who develops hypotheses also becomes responsible for screening respondents and cleaning data. As shown in Fig. 1, compared to actively-managed services, the use of minimally-managed services is growing quickly because of the lower cost perceptions and greater convenience (Shapiro, Chandler, & Mueller, 2013).

The typical price paid for an MTurk survey response varies between $0.50 and $1.50. Data obtained through Qualtrics is normally approximately $5 dollars per "good complete" (complete responses surviving screening by data provider) for household data. That charge grows for more specialized panels. For example, a panel of business-to-business (B2B) salespeople may cost $20 or more per good complete. Given the economics and convenience, it is not surprising that academics have embraced the use of MTurk (Hulland & Miller, 2018).

[1] Amazon more recently uses the term project for a HIT. According to the MTurk website, achievement of the "Master" designation is determined by statistical models. The MTurk worker blogs indicate that thousands of HITs must be completed in order to meet statistical criteria.

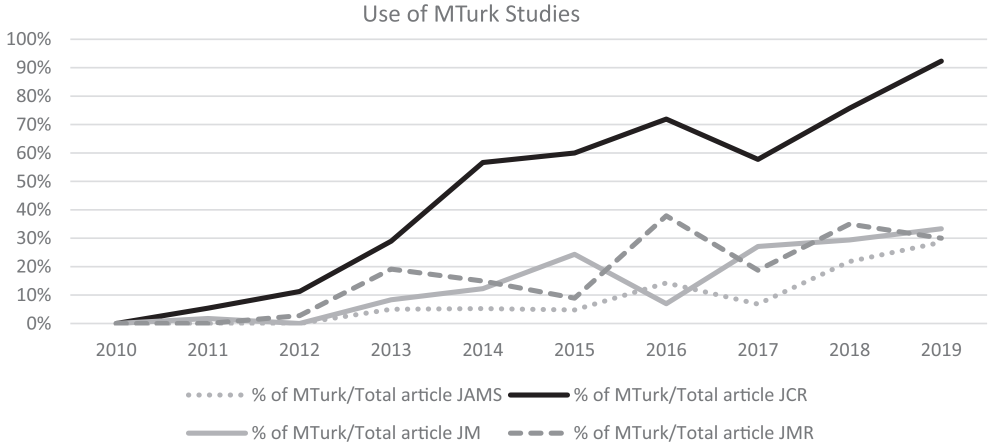

Fig. 1. Prevalence of studies that used at least one MTurk sample.

2.1. Who are minimally-managed panel respondents?

Although the crowdsourcing platform Prolific is gaining popularity, the most commonly used minimally-managed online data collection service is MTurk. Amazon provides no official information regarding the total size and aggregate demographics of MTurk "workers." Information on just who MTurk workers are remains scarce. Paolacci and Chandler (2014) stated that there were more than half a million MTurkers living in 190 countries in 2014, which is consistent with Amazon's advertised size. According to Robinson, Rosenzweig, Moss, and Litman (2019), there are about 250,810 active workers, of which there are 4845 "superworkers" who account for nearly half of the total HITs completed. Superworkers are likely over-sensitized to research materials, particularly experimental manipulations, because they participate in so many studies. Because of over-sensitization, one could make the argument that superworkers are quite different in their perceptions from less sensitized participants, again raising serious questions about representativeness.

Further complicating the effort to identify respondents, Virtual Private Networks (VPNs) can disguise the true physical location of respondents. Dennis, Goodson, and Pearson (2018) directly investigate the use of VPNs in MTurk samples and identify that about 30% of respondents use VPNs. About 85% of observations associated with VPNs include nonsensical responses that the authors suspect as caused by a lack of English skills (Dennis et al., 2018), calling into question whether respondents are truly located in the U.S. While Prolific does provide a mechanism for participants to verify their identity, most minimally-managed crowdsourcing participants are essentially anonymous.

A major concern with samples from minimally-managed online services is repeated participation, where respondents take part in the same or similar studies many times (Woo, Keith, & Thornton, 2015). Using VPNs, a single respondent may complete the same survey multiple times with multiple different accounts (Chmielewski & Kucker, 2020; Chesney & Penny, 2013). Even without VPNs, MTurkers may complete separate tasks for the same requester without restriction (Cheung, Burns, Sinclair, & Sliter, 2017). With repeated exposure to the same or similar study materials, respondents can learn a study purpose, expose themselves to information from other experimental conditions, and ultimately determine the response that the researcher is looking for (Conte, Levati, & Montinari, 2014).

2.2. Data quality

Overall, the attentiveness and scale reliabilities of samples collected from minimally-managed services appear to be comparable or exceed those obtained from actively-managed services. Kees et al. (2017) compare a series of different samples obtained through Qualtrics with MTurk and other sources like student samples and Lightspeed. There were some issues with all sources, but they conclude that the data from MTurk was no worse than any other potential sources. In fact, they argue that MTurk data actually outperformed professional panels based on participant involvement, attention checks, research participation and general computer knowledge. Yet, other authors recently point to "an MTurk Crisis" based on the declining data quality evidenced recently (Chmielewski & Kucker, 2020).

The marketing literature displays scant attention to data screening. As shown in Table 1, only about 16% of articles using MTurk samples discuss screening procedures. Even when mentioned, articles often only provide minimal information about screening, such as country restrictions (e.g., "a U.S. sample") or approval rating (e.g., "1000 HITs with a 95% approval rating"). Without a systematic approach to understanding screening criteria, it is impossible to estimate data quality. In regards to screening, Chandler, Mueller, and Paolacci (2014, p. 121) argue, "What is particularly worrying is that even a casual inspection of the papers that use data collected from MTurk reveals that workers are frequently excluded for a wide variety of reasons and that these exclusion criteria are often applied post hoc."

Table 1
Reports of MTurk screeners.

	JAMS	JCR	JM	JMR	total
MTurk articles	29	264	54	87	434
Reports screeners	5	38	12	15	70
Percentage reporting	17%	14%	22%	17%	16%

Screening can occur prior to data collection and prevent unsuitable respondents from participating; or screening can be post-hoc, in which unsuitable observations are removed. Unlike actively-managed online crowdsourcing services, any willing member of a minimally-managed online crowdsourcing service can participate in a study unless barred by a screening question. If researchers do not carefully select appropriate pre-screening questions, it is likely that many nontargeted respondents will participate. Babin, Griffin, and Hair (2016) suggest that authors' data cleaning processes should be described thoroughly as an important consideration to ensure that data cleaning is done carefully and legitimately. The concern is screening that eliminates data points without proper justification (Ford, 2017). Post-hoc screening provides an opportunity, consciously or unconsciously, for researchers to screen observations based on finding support for whatever hypotheses might be proposed (Chandler et al., 2014). Thus, there is a need to develop a framework for applying appropriate screeners to reduce the number of nontargeted respondents from participating and to ethically clean data. The academic marketing literature has yet to delineate a series of screening mechanisms for researchers to follow.

DeSimone and Harms (2018) describe the general state of survey data screening on data quality and categorize screening tools into three types: direct, unobtrusive, and statistical. Direct methods of screening involve standard attention filter questions. Unobtrusive screening examines patterns of responses and screens out those with implausible response patterns. Statistical techniques like long-string tests can identify low-quality responses. Perhaps most disturbing, the research suggests that employing all types of screening would lead to deletion of more than 80% of responses in typical online surveys. Thus, adding to whatever percentage of sampling units not replying, missing data rates could be extraordinarily high. As the authors recommend, "due to the small amount of overlap between screening indices, researchers should not hesitate to employ multiple screening techniques to identify LQD (low-quality data) in their data" (DeSimone & Harms, 2018, p. 575).

3. Screener typology

A valid sample should accurately represent a specific and intended population. For data involving human participation, respondents should match the target population and should provide thoughtful and truthful answers (Berinsky et al., 2012). Consistent with the goal of producing representative and accurate samples, screening criteria that fall into two primary categories should be deployed: (1) *Selection screeners*, used to identify respondents who truly fit a sample profile that matches the relevant target population, and (2) *Accuracy screeners*, which identify whether respondents are providing truthful and appropriately thoughtful answers to questions. The methodological categories of screeners presented by DeSimone et al. (2015) can be further condensed into two categories: (1) *Direct screeners*, which relate to how respondents answer screening questions, and (2) *Statistical screeners*, which relate to the use of analytical methods for screening respondents. In total, we recommend a two-by-two framework for screening, shown in Table 2:

3.1. Screener goals

Screeners can have the goal of improving selection or accuracy. Selection screening criteria are used to identify respondents who fit relevant target population characteristics. Selection screeners relate to sampling coverage bias, which is whether the sample represents the entire population and no more (Fricker, 2008). Whereas researchers using traditional sampling methods were almost entirely concerned with overrepresenting certain segments of a population and missing/underrepresenting other segments, coverage bias for online surveys also includes inadvertently sampling nontargeted populations. Because crowdsourced respondents are paid based on the number of requests completed, they are incentivized to participate in as many surveys as possible regardless of whether they match the profile of the targeted population. Crowdsourced surveys can be accessed from any country, and the small payments that seem trivial in the U.S. can be extremely desirable in other countries. Using comparative salary data from 2014 (the latest information available), a fifty-cent payment for a survey in the U.S. would be equivalently valued at about $2.25 in China and $4.90 in the Philippines (NationMaster, 2019). Consequently, people from poorer countries have more incentive to participate than people from wealthier countries.

Selection screeners are essential for research involving online data collection. Even when theory should generalize broadly to humans worldwide, research materials must still be adapted to a limited set of languages/cultures/countries. For example, Woolley and Risen (2021) used scenario-based experiments with MTurk samples from the U.S. Although their theory should be widely generalizable to people around the world, their research materials were customized to the language and culture of the U.S., including presenting respondents with an American menu that included calorie information. Respondents from Asian countries may view calorie information and American food choices very differently from American respondents. Thus, sample representativeness is still important for widely generalizable theory because of the need to customize

Table 2
Example screeners by purpose and methodology.

	Methodology	
Purpose	Direct	Statistical
Selection screeners	MTurk-provided, Captcha verification, IP-address match, Self-report state, ZIP-code match, State capital quiz, honesty trap.	Motorcycle ownership, Demographics
Accuracy screeners	MTurk-provided, pick "2", Attention prediction quiz (oak tree test), open-ended responses (race and income)	Time spent completing the survey, straight-lining, personal reliability

research materials. The degree to which sampling bias is considered acceptable depends on the research purpose, but researchers should at least be aware of potential problems. Selection screeners provide a mechanism for researchers to improve, or at least estimate, sample representativeness.

Accuracy screeners, on the other hand, attempt to identify whether respondents are providing truthful and appropriately thoughtful answers to questions. Accuracy screeners include attention checks, quality predictors, and response quality. Attention checks determine whether respondents are attentive to research material (Paolacci, Chandler, & Ipeirotis, 2010). Subject inattentiveness occurs when participants do not follow the instructions or carelessly answer questions (Hauser, Paolacci, & Chandler, 2019). Because some online data collection services incentivize participants by survey, survey takers are motivated to complete as many surveys as possible; hence, subject mindlessness is a significant concern (Ford, 2017). Indeed, MTurk workers admit to multitasking (Chandler et al., 2014) and tend to complete surveys faster than other populations (Kees et al., 2017; Smith, Roster, Golden, & Albaum, 2016). Inattentive responses increase measurement error in small samples and systematic error in large samples (Hauser et al., 2019). For example, respondents selecting the same answer for all measurement questions can lead to spurious correlations, inflate internal reliability (Wood, Harms, Lowman, & DeSimone, 2017) and decrease discriminant validity (Hamby & Taylor, 2016).

3.2. Screener methodology

Screener methodology can be either direct or statistical. Direct screeners include any item and methodology with a primary purpose of identifying problematic responses and does not rely on overall sample statistics or causal analysis of focal variables. Direct screeners have the advantage of clearly identifying problematic responses but the disadvantage of being visible to respondents. Examples include attention check questions and response matching (i.e., asking the same question multiple times) on control variables. A key strength of direct screeners is that researchers can use them to *clean* data prior to any substantive data analysis. Thus, it is easier to make unbiased decisions about data screening prior to any substantive examination.

Statistical screeners estimate sample quality using analyses that are typically undetectable to respondents. The categories of "unobtrusive" and "statistical" screeners are combined because they are both unobtrusive and rely on statistical calculations to estimate response quality after data has been collected (DeSimone & Harms, 2018). They were originally separated "due to the computational effort involved" (DeSimone & Harms, 2018, p. 561), but computational effort is not relevant for the purpose of understanding the benefits and limitations of screeners. Whereas direct screeners can theoretically be used to screen respondents at any point, including while respondents are taking the survey, statistical screeners can only be used after data has been collected. Examples include comparing a sample statistic to a known norm, duration of time spent taking the survey, outlier analyses (e.g., Mahalanobis distance), psychometric synonym-antonyms (conceptually opposing terms that should produce opposite responses), and spotting straight-lining behavior (DeSimone et al., 2015).

3.3. Direct selection screeners

Direct selection screeners filter out ineligible respondents by attempting to prevent participation from individuals who do not belong to the targeted population. Direct selection screeners filter based upon sample attributes, such as geographic location (e.g., country of residence: U.S.), demographic characteristics (e.g., gender: female), socio-economic factors (e.g., education level: college degree), job type (e.g., industry: professional selling), physical or psychological traits (e.g., body modifications: tattoos), psychographic characteristics (e.g., hobbies: travelling), and ownership (e.g., motorcycle). For example, if a researcher is interested in learning about the employee compensation plans of professional salespeople, then direct selection screeners should filter out respondents not employed in professional selling.

Typically, researchers assess direct selection screeners using simple straightforward measures. For example, to assess industry, respondents are typically given a list of industries and asked to select the one in which they are currently working, which is fairly easy to manipulate. Rand (2012) used IP addresses to verify MTurker self-reported location and found a 97% matching rate. However, the use of MTurk outside of the U.S. has grown dramatically since that study, and location screeners are complicated by the use of VPNs. To augment simple measurements, researchers should consider including additional direct-selection screeners. For location, researchers might use direct screeners quizzing respondents about postal codes or state capitals or regional landmarks in addition to reliance on self-report location or IP-address. Then, the additional questions can be matched to IP-address providing the researcher with the opportunity to estimate the likelihood the respondent belongs to the target sample location.

The most common screener for detecting multiple responses is IP-address; however, as mentioned previously, VPNs can nullify screening for duplicate IP-addresses. Another screener is an open-ended question asking for the study purpose used in conjuncture with a between-subjects design. Identical answers to open-ended questions can be a sign of multiple responses. Also, if respondents provide a response that assumes knowledge of multiple conditions, it is an indication of taking the survey multiple times. As for identifying sub-segments of a population, screeners can be designed to look for specific characteristics such as a type of job, a demographic characteristic, or a type of behavior/personality trait.

To further determine honesty about identity, researchers can ask information multiple times or in multiple ways (Huff & Tingley, 2015). Wessling et al. (2017) used this method to demonstrate that a large portion of MTurk respondents misrepresent themselves. They used a multipart study in which data were collected at one time period and then again shortly thereafter. Interestingly, 17% of an MTurk sample screened as smokers, over 50 years old, and being treated for lung cancer, also participated in (as matched by worker ID) a study of active athletes under 35 years of age being treated for a dislocated shoulder. To try to screen for dishonest self-descriptions, surveyors can ask about identifying attributes twice (or more) in a single survey. Alternately, the respondent can be asked about an attribute using a multiple-choice question and then asked to provide more detail about it using an open-ended question. Asking the question several times using the same wording should probably be avoided as the respondent may be sensitized to look for certain types of questions. Furthermore, adding questions can increase fatigue and irritation, which can lower the quality of the other responses.

3.4. Statistical selection screeners

Statistical selection screeners assess whether the data is likely to be representative of the target population. Statistical selection screeners are used in three steps. First, researchers find a characteristic that is known for a target population and several nontargeted populations. Second, researchers ask respondents about that characteristic on the survey. Finally, that information is then compiled for the sample and compared to the target and nontargeted populations. If there is a statistical difference between the sample and the target population, then the sample would contain a significant percentage of respondents who are not representative of the target population. If the sample matches a nontargeted population, the sample is likely to be comprised of a large percentage of respondents from that population. To screen for country of origin, example characteristics could include motorcycle ownership, the number of bedrooms/bathrooms in respondents' home or whether respondents ate pizza in the previous month.

For some research purposes, demographic characteristics may be different between targeted and nontargeted populations. For example, if the gender composition of a target population is known, then a sample with a significantly different gender composition could indicate non-representativeness. If gender impacts response patterns or correlations among variables, then hypothesis testing could be impacted if researchers do not statistically account for the gender mismatch. Furthermore, a mismatch between a sample and target population on a particular variable could indicate that the sample either over- or under-represents the target population. If the findings from the target population are expected to be different than other populations, then a sample mismatch would be problematic for hypothesis testing. For example, if a researcher has designed research materials to be understood by a North American sample of English speakers, and the sample statistics do not match the known characteristics of that group, then a large percentage of respondents may not respond to the research materials as anticipated, providing potentially inaccurate and misleading results.

3.5. Direct accuracy screeners

Direct accuracy screeners assess attention and response quality. The most straightforward direct accuracy attention check screeners are questions quizzing respondents about information directly shown in the research materials. For example, respondents might answer a quiz about an experimental scenario, or be asked to select "2" or "disagree," as a response on a Likert-type scale. Respondents providing incorrect answers may then be removed. Another method is to ask open-ended questions and remove respondents who provide nonsensical responses. Interestingly, Downs, Holbrook, and Peel (2012) found that participants who failed attention check questions did not perform any worse in reliabilities than those who passed. Yet, this might depend on how carefully the attention checks were worded, and whether people were being sensitized about what to watch for (e.g., from online discussion forums or prior experience).

Several other techniques can predict response quality to research materials. One technique is to have respondents take a quiz on material unrelated to the research purpose so that respondents can be eliminated prior to data collection. One version of this screener is the "bamboo" test in which respondents are asked to read a paragraph about bamboo and take a quiz on it. In the paragraph, respondents are instructed to select "two" for each quiz question.[2] Another way to predict response quality is to use past performance. MTurk has provided screeners for specifying the number of HITs Turkers have completed, the percentage of HITs that have been approved for payment by other requesters, and whether Turkers have become "Master"

[2] We became aware of the bamboo test via word-of-mouth from other research teams, and we could not find a formal reference to it in the academic literature.

workers, indicating that they produce high quality work (Matherly, 2019). Master workers are tantamount to professional survey takers. Compared to undergraduate student subjects, master workers complete dozens, if not hundreds, of times more tasks than typical student respondents. Thus, their experience makes them abnormally sensitive to elements of the design and much more sensitive to demand characteristics (Babin et al., 2016; Smith et al., 2016).

3.6. Statistical accuracy screeners

Statistical accuracy screeners involve the use of methodologies that are "undetectable" to respondents, such as response times and response patterns to assess attention and honesty. Examples include calculating response time compared to the sample mean to detect "speeders" and "slow-pokes," psychometric antonyms, and Mahalanobis distance (DeSimone et al., 2015). Because these screeners are "invisible" to respondents, it is difficult for respondents to falsify attention or otherwise "game" statistical accuracy screeners. As such, statistical accuracy screeners provide important evidence of data quality.

However, it is more difficult to set objective thresholds for removing data when population statistics are unknown, which is the case for most statistical accuracy screeners. Outlier analyses can be used to identify problematic cases, but it is incumbent upon the researcher to select the appropriate test for a given set of data (Aggarwal, 2017; Woolrich, 2008). When data are normally distributed and have a distribution that approximates the target population, researchers can use standard scores (Aggarwal, 2017). Hair, Black, Babin, and Anderson (2019) recommend a general heuristic of 2.5 standard scores for small data sets ($n < 81$) or 3–4 for larger samples. When these assumptions do not hold, methods such as probabilistic models (e.g., Gaussian mixture models) and proximity-based models (e.g., density-based methods) can be used.

Selecting the appropriate outlier analysis is not always straightforward. Consider the use of survey completion time as a screening criterion for removing responses. "Speeders" are respondents who rush through the survey and who are presumably not paying sufficient attention (Smith et al., 2016); "slow-pokes" take too long to complete the research materials. Completion time would have a chi-square distribution rather than a normal distribution and thus would be ineligible for the standard scores analysis. Furthermore, if speeding was widespread in a sample, then that sample's distribution would establish an overly conservative threshold for identifying "speeders" when compared to a sample with relatively few speeders. This is problematic because it is believed that MTurk samples contain a large proportion of speeders (Ford, 2017). If so, then researchers would set less rigorous standards for identifying speeding for MTurk samples. For these reasons, researchers should not use standard scores for detecting problematic responses based on completion time.

Huang, Curran, Keeney, Poposki, and DeShon (2012) present research suggesting that an index of less than 2.0 sec/survey item is indicative of a speeder (DeSimone & Harms, 2018). Additionally, if an experiment uses an audio–video clip as a manipulation, then respondents must watch the audio–video clip, in order to have participated in the experiment. If respondents spend less time engaged with the clip than its objective length (e.g., less than one minute on a one-minute clip), then they could not have been attentive to the entire clip and should be removed. While this method cannot identify which respondents were attentive, it can easily identify respondents who were not.

Similarly, for attentiveness to the survey, respondents can be excluded if they provide the same response for a large number of consecutive items or synonym-antonym items with opposing descriptors (i.e., bad-good, favorable-unfavorable). Thus, it is possible to set relatively objective thresholds for removing data using statistical accuracy screeners, as long as researchers use appropriate criteria.

4. Methodology

The goal of the methodology is to compare each of the four categories of screeners (i.e., direct selection, direct accuracy, statistical selection, and statistical accuracy) by data source. The findings are not intended to be generalizable, but they will provide insight into how screening procedures impact sample size by data source.

4.1. Samples

The research questions were tested using an online survey. The stated purpose of the survey was how memory and distraction influence first impressions of service providers and comprehension of marketing messages. Additionally, respondents were informed that the research was to involve Americans living in the Southern United States. "Please only complete this survey if you live in the Southern part of the U.S., which includes the following states: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia." Samples were drawn from Dynata (formerly SSI), Qualtrics panel aggregation, three distinct MTurk postings with different built-in accuracy screeners, and an online university subject pool. Qualtrics and Dynata, as actively-managed data collection services, were selected because both have strong reputations, and MTurk was selected because it is the most popular minimally-managed online data collection service. A student sample was also chosen as a basis of comparison.

All samples were compensated using the normative standard payment for each data set (50 cents for the less screened MTurk samples, $1.50 for the master MTurk sample, $5 per "good complete" survey for Qualtrics and $3.75 per "good complete" survey for Dynata[3]) except for the student sample, which had a choice of studies to complete for class grade points. The price paid by the researchers is not necessarily the payment received by the respondent. Whereas the MTurk sample payment represents the exact amount paid to respondents, the other sample payments reflect the amount paid to the service, of which a portion is allocated to the respondent. Services may also offer additional incentives not reflected in the total payment, such as respondent quality indicators. Thus, it is difficult to compare the exact compensation received by each respondent across samples. If compensation influences attention, then this could be a confounding issue. On the other hand, it is important to pay respondents at the typical expected rate; otherwise, respondents may answer the survey abnormally. The key attraction of MTurk is the low cost to researchers. The purpose of this research is to compare "typical" samples. If some respondents were paid an atypically high rate while other workers were paid at a typical rate, the results may be misleading. As such, payment was set based on the "going rate" for each sample to attract typical participants for each service.

Each panel was to include 200 respondents; however, Dynata charges a minimum fee, so they provided 418 responses, and 154 Master MTurkers and 134 students completed surveys. Project managers at Dynata and Qualtrics will typically remove responses that they deem insufficient quality as a service; however, the project managers in this research were instructed not to remove respondents (except incomplete responses) so we could assess the quality of the attention check questions.

4.2. Manipulated conditions

Multiple forums exist through which MTurk workers discuss HIT requests, including how one should respond to get paid. Because of the nature of the present study, the concern was that participants would recognize the study purpose and inform one another. To determine whether respondents were meta-gaming the survey, a "red-herring" between-subjects experimental design was utilized in which respondents were shown either a black or white service provider. With a between-group experimental manipulation, respondents will only see one of the manipulations. If the respondent is aware of both manipulations, it would mean that the respondent had either taken the survey multiple times or communicated with another respondent. Additionally, the items for perceived similarity were either grouped together in a single block of questions or mixed-up with other questions related to service-provider benevolence. Varying the presentation should affect respondents more if they are inattentive and more prone to bias by the survey format. If there is a significant difference in response variation to perceived similarity between the mixed and together conditions, it means the respondents are susceptible to the survey format.

4.3. Selection screeners

Selection screening aims to verify that respondents individually are members of, and the sample overall, is representative of the intended target population. In this particular survey, the target population is adult consumers from the southern U.S. Consequently, many of the selection screeners are related to location. For supplier-provided filters, the actively-managed services, Dynata and Qualtrics, provide their own external screening mechanisms, and true scientific panels draw samples using accurate sampling frames that prevent mistaken identities. MTurk provides automatic screeners as an option, and in this case, U.S. respondents only were requested. However, no premade mechanism was set up to prevent people from non-Southern states from participating so that it would be possible to ascertain the effectiveness of using written instructions to accurately limit eligibility. As a final supplier-provided selection screener, a captcha verification provided by Qualtrics was included at the end of the survey. Captcha verifications are challenge-response questions based on visual information, such as "click on each of the following pictures that contains a stop sign" or "what letters do you see?" Captcha verifications are theoretically difficult for automated "bots" to complete, so they are supposed to provide a barrier to automated responses. Captcha verifications are intended to screen out respondents who are not part of the sample (i.e., bots), so they are considered a type of external selection screener. Qualtrics provides two mechanisms for using the question. Either respondents must successfully answer the captcha verification to proceed or a proprietary algorithm is used to estimate the likelihood the response is from a bot. Rather than rely on the algorithm, we chose to include the captcha verification question on the penultimate page of the survey. If a survey were stopped at approximately 95% completion, it would indicate that the respondent failed the captcha verification.

For user-generated, direct-selection, screeners, location was assessed using IP-address (automatically captured) and self-reported ZIP code and state of residence. Additionally, the survey asked each respondent to report the capital city for their state. While it is possible to look up information about ZIP codes and state capitals, it is unlikely that respondents will take the time to do so as it will slow them down and reduce the number of surveys they can complete. In addition to location, a lie trap was included to check whether respondents were truthful about their gender. The lie trap asked respondents their gender early in the survey. Much later in the survey, respondents were told, "Please complete the last section only if you are ___

[3] Panel service firms often perform some direct screening before providing data to the analyst. Routinely, the services flag respondents not completing the survey, failing direct attention checks, or completing the task too quickly (determined by their own criteria). A "good complete" then passed direct screeners and did not speed. In those cases, customers pay only for good complete responses.

{Unselected Gender}. Respondents completing the last section may be eligible for a small bonus. What is your gender?" Thus, the question automatically prompted the opposite gender. The lie trap is similar to one used by Wessling et al. (2017), except that their study collected the same variable in separate surveys given at different times instead of within the same survey.

We also included several selection sampling quality screeners. First, we asked respondents whether they owned a motorcycle and then compared the percentage ownership rate to documented ownership. In the U.S., motorcycle ownership in the Southern U.S. was calculated by dividing the total number of registered motorcycles in the Southern U.S. for 2019 by the adult population in those states, which was approximately 3%. By contrast, some estimates show that 80% of families have a scooter or motorcycle in Southeast Asia and 60% have one in China (Pew Research Center, 2015). Hence, because motorcycle ownership is relatively low in the U.S., we can estimate the likelihood our sample is from the U.S. (or from a given state as statistics are available by state) based on motorcycle ownership as a benchmark. Second, we can compare the demographic characteristic of respondents in each sample to the known demographic characteristics of the Southern U.S. population to look for discrepancies.

4.4. Accuracy screeners

Three different MTurk samples were selected: one with no accuracy screener options (MT0), one with low accuracy screener options (MTL), and one with high accuracy screener options (MTM). In addition, the selection screener requiring that respondents must be from the U.S. was applied to all samples. The MT0 sample used no other supplier-provided screener. The MTL sample required worker accounts to have completed at least 50 HITs with a minimum of a 95% approval rating. The MTM sample required workers with a "Masters" designation, at least 1,000 HITs, and a 95% approval rating. Researcher-generated, direct-accuracy screeners included a pick "2" question, logical responses to open-ended questions about race and income, and a quality prediction quiz on an informational paragraph about oak trees. The typical bamboo quiz was not used because the assessment is discussed in MTurk forums (reddit.com/r/MTurk), which could compromise its usefulness. While "workers" are not always discussing what they encounter, researchers need to avoid using the same screeners in all surveys.

Accuracy statistical screeners include time spent taking the survey, straight-lining behavior, personal reliability, and reliability between survey formats. First, mean and range time spent on the survey were calculated. According to Qualtrics, the survey should take approximately 13 min to complete. Respondent taking less than 20% of 13 min (<156 s) were marked as "speeders" and those taking 80% longer than 13 min (>1404 s) were marked as "slow." DeSimone et al. (2015) lists a number of screening techniques that require data analysis. To address those screeners, scales related to similarity-attraction and trust were included on the survey. Respondents were shown a picture of a service provider (selected at random) and were asked to rate the service provider on perceived similarity and trust based on their first impression. Perceived similarity is the idea that one person feels that they share attitudes, opinions, beliefs, and characteristics with another person and is not necessarily related to actual similarity (Montoya, Horton, & Kirchner, 2008). Perceived similarity was measured here using a five-item, five-point Likert-type scale (Feick & Higie, 1992). Expertise was measured using a five-item scale, and benevolence was measured using a four-item, five-point Likert-type scale (Wood, Boles, Johnston, & Bellenger, 2008). Trust was measured using a single bipolar adjective scale that ranges from −3 = Very untrustworthy to 3 = Very trustworthy. Per the procedure by DeSimone et al. (2015), personal reliability was calculated for perceived similarity by splitting the items into two separate scales and correlating them using Spearman's Rho. Additionally, the variation in response between format condition was assessed, as was the practice of "straight-lining," which was evaluated by identifying and counting respondents who used the same anchor for every scale item of similarity, expertise, and benevolence, such as all "4's" or all "7's."

5. Results

5.1. Direct selection screeners

Direct selection screeners assess the likelihood that each respondent is a member of the desired target population. In this study, the target population includes human respondents in the Southern US. Additionally, direct selection screeners were used to assess respondent honesty about geographic location and demographics. The results are shown for each sample in Table 3.

Qualtrics captcha verification was used to eliminate automated respondents (non-humans "bots"). However, it eliminated only a single response, and because the response was from the student sample, it could not be a bot. That either means none of the responses in any of the data sets were bots or the captcha verification question was ineffective at removing bots. Related to geographic location, the MTurk samples with the lowest premade screeners (i.e., MT0 and MTL) were the least likely to be located in the Southern US. The masters MTurk sample were more likely to report a Southern US state than other MTurk samples but less than the actively-managed samples. Related to IP-address, the MT0 and MTL samples had fewer USA IP addresses and the match between IP address and stated location was lower. Furthermore, they were much less likely to indicate residing in the Southern US, suggesting that either the MT0 and MTL samples were inattentive to the survey directions or hoped to participate regardless of their state of residence.

Table 3
Direct selection screener findings.

Researcher added Screeners	MTO[a]		MTL[a]		MTM[a]		Qualtrics		Dynata		Students	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
None	200	100%	201	100%	154	100%	212	100%	418	100%	134	100%
Captcha	200	100%	201	100%	154	100%	212	100%	418	100%	133	99%
IP USA	185	92.5%	186	92.5%	152	98.7%	211	99.5%	415	99.3%	134	100%
IP Matches	88	44.0%	103	51.2%	131	85.1%	174	82.1%	346	82.8%	126	94.0%
Self-report state	89	44.5%	92	45.8%	98	65.3%	212	100.0%	408	97.6%	129	97.0%
ZIP code matches	166	83.0%	178	88.6%	138	89.6%	208	98.1%	412	98.6%	128	95.5%
State Capital Quiz	105	52.5%	123	61.2%	127	82.5%	176	83.0%	352	84.2%	122	91.0%
Gender Honesty	158	79.0%	170	84.6%	134	96.4%	195	92.4%	396	94.7%	129	97.0%
Passed all	34	17.0%	42	20.9%	70	45.5%	140	66.0%	288	68.9%	112	83.6%

[a] MTurk samples: MT0 = no accuracy screeners, MTL = low accuracy screeners, MTM = high accuracy screeners.

Finally, honesty was evaluated by comparing stated location to ZIP code, asking respondents to name their state capital, and the lie trap for gender. Consistent with the other findings, honesty was lowest among the MT0 and MTL samples, particularly for the state capital quiz and the lie trap. The ZIP code match was lowest for the MT0 sample and approximately equivalent for the MTL and master MTurk sample. The master MTurk sample had one of the highest honesty rates for the lie trap, however, indicating that they tend to report consistent demographic characteristics to received additional compensation.

In sum, only 17% of MT0 respondents passed all of the direct selection screeners compared to 68.9% of the Dynata sample and 83.6% of the student sample. In this research, the actively-managed samples were more likely to represent the target population than the MTurk samples, with the noticeable exception that the master MTurk sample were among the most consistent about their gender.

5.2. Statistical selection screeners

The demographic characteristics of each sample were compared to the population to approximate the representativeness of each sample. The results are presented in Table 4. The adult population of the Southern US is slightly skewed toward female. The MT0 (χ^2 = 19.25, p < .001), MTL (χ^2 = 3.12, p < .1), and Qualtrics (χ^2 = 48.69, p < .001) samples all had a statistically different gender composition than the population. The MT0 (χ^2 = 6.13, p < .05), MTM (χ^2 = 12.76, p < .001), and Dynata (χ^2 = 6.78, p < .01) samples consisted of more people identifying as white than the population, while the student sample had fewer people identifying as white than the population (χ^2 = 15.44, p < .001). Respondents in all of the samples MT0 (χ^2 = 152.06, p < .001), MTL (χ^2 = 118.83, p < .001), MTM (χ^2 = 21.17, p < .001), Qualtrics (χ^2 = 3.052, p < .1) and Dynata (χ^2 = 3.93, p < .05) were more likely to have a college degree except for the student sample, who typically had yet to earn a degree. The MT0 (t = –9.92, p < .001), MTL (t = –7.40, p < .001), and student (t = –25.61, p < .001) samples were all younger than the population while the Qualtrics (t = 3.73, p < .001) was older. The MTM (t = –1.90, p < .1) sample reported a lower income while the student (t = –2.41, p < .05) sample reported a *higher* income. Finally, motorcycle ownership was higher than the population for every sample that included data: MT0 (χ^2 = 2155.33, p < .001), MTL (χ^2 = 1542.00, p < .001), MTM (χ^2 = 20.97, p < .001), Qualtrics (χ^2 = 62.53, p < .00) and Dynata (χ^2 = 134.58, p < .01).

Additionally, there are a number of outliers and anomalies in the data. Race, age income for each respondent were measured using open-ended questions for the purpose of providing additional insight into response quality. For race, 7% of the MT0 and 8% of the MTL samples provided answers not related to race, such as "Female," "Ghost," "Babyboy," "RACE," and "dsfsdfwe," compared to only 3% of Dynata respondents, 2% of student and Qualtrics respondents, and less than 1% of MTM respondents. Related to age and income, there were a few outliers, particularly in the Dynata sample.

In sum, none of the samples represented the target population demographics on all of the measured characteristics. That said, the MT0 sample demographic characteristics was the poorest statistical match with the target population, only matching the target population on stated income.

5.3. Direct accuracy screeners

The direct accuracy screeners evaluate attention and data quality. The results for each accuracy screener are present for each sample in Table 5. The direct accuracy screeners included the "pick 2" question, the quality prediction quiz, and the aggregation of the open-ended responses from the demographic questions. Based on the results of a one-way ANOVA with a Tukey post hoc test, the MTM sample was more likely to pass the "pick 2" screener than any other sample (f (5, 1294) = 8.7, p < .001, with all mean differences significant at p < .05 or better). The MTL sample was also more likely to pass the screener than students. The other samples passed at comparable rates. The MTM sample was also more likely to pass the quality prediction screener (f (5, 1294) = 7.39, p < .001, with all mean differences significant at p < .05 or better). As before, the other samples passed at comparable rates. Finally, the open-ended questions were assessed using a cross-tabulation of observed

Table 4
Statistical selection screener findings.

		MT0[a]	MTL[a]	MTM[a]	Qualtrics	Dynata	Students	Pop est.[b]
	Sample *n*	200	201	154	212	418	134	
Sex	Male	64%	54%	51%	24%	45%	51%	48%
	Female	37%	46%	49%	76%	55%	49%	52%
Race	White	72%	66%	78%	65%	68%	46%	62%
	Black	9%	8%	9%	22%	14%	29%	21%
	Asian	7%	7%	7%	5%	6%	10%	3%
	Hispanic	4%	9%	3%	2%	6%	7%	12%
	Other	2%	4%	3%	4%	3%	6%	3%
	Not a race	7%	8%	0%	2%	3%	2%	–
Education	>HS	0%	0%	0%	2%	3%	0%	11%
	HS	14%	20%	25%	47%	41%	41%	46%
	Assoc.	12%	10%	24%	24%	19%	47%	10%
	Bach.	57%	52%	37%	20%	24%	11%	21%
	Mast.	15%	16%	11%	6%	10%	2%	10%
	Doc.	3%	0%	2%	1%	4%	0%	2%
Age[c]	Mean	32.4	33.5	40.5	43.3	139.8[b]	24.2	39
	Std. Dev	9.4	10.5	26.4	16.9	1964.8	6.6	–
	Min	18	18	24	12	5	10	–
	Max	68	70	333	80	40,213	52	–
Income[c]	Mean	69k	70k	57k	49k	312k	83k	56k
	Std. Dev	186k	127k	47k	48k	5003k	120k	–
	Min	0	0	0	0	0	0	–
	Max	2500k	1300k	365k	300k	100,000k	1000k	–
Motorcycle ownership	Yes	59.0%	50.2%	9.1%	12.3%	12.7%	NA	3%

[a] MTurk samples: MT0 = no accuracy screeners, MTL = low accuracy screeners, MTM = high accuracy screeners.
[b] Population estimates for age, race, and income for the Southern U.S. states were drawn from the Kaiser Family Foundation (kff.org) for 2019. Educational attainment was attained from the U.S. census for the entire United States 2019 to reflect the most accurate data available. Motorcycle ownership was drawn for motorcycle registrations for the Southern U.S.
[c] No data were removed or cleaned prior to reporting these results. As a result, some data may be abnormal.

Table 5
Direct accuracy screener findings.

Researcher added Screeners	MT0[a]		MTL[a]		MTM[a]		Qualtrics		Dynata		Students	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
None	200	100%	201	100%	154	100%	212	100%	418	100%	134	100%
Pick “2”	175	88%	177	88%	139	90%	179	84%	335	80%	102	76%
Pred. quiz	88	44%	114	57%	101	66%	101	48%	207	50%	64	48%
Open-ended	185	93%	187	93%	150	97%	203	96%	406	97%	125	93%
Passed all	78	39%	102	51%	101	66%	91	43%	184	44%	54	40%

[a] MTurk samples: MT0 = no accuracy screeners, MTL = low accuracy screeners, MTM = high accuracy screeners.

versus expected poor-quality answers to open-ended questions. The finding shows that the MT0 and MTL samples observed more poor-quality answers than expected as compared to the other samples, which all had fewer ($\chi^2 = 11.56$, df = 5, $p < .05$).

In sum, the results show that respondents in the MTM sample were more likely to pass direct accuracy screeners than in the other samples. However, respondents in the MT0 and MTL samples were more likely to give nonsensical answers to open-ended questions. Thus, there is a discrepancy among the MTurk samples related to passing direct accuracy screeners.

5.4. Statistical accuracy screeners

The statistical accuracy screeners provide insight into the quality of responses provided by respondents. The results are provided in Table 6. In regard to time spent taking the survey, there was no statistical difference between the number of respondents going more slowly than expected ($\chi^2 = 2.46$, df = 5, ns). However, all of the MTurk samples contained more speeders than expected ($\chi^2 = 23.45$, df = 5, $p < .001$), while the other samples contained fewer speeders than expected. Conversely, related to straight-lining, all of the MTurk samples and the Qualtrics sample contained fewer respondents engaging in straight-lining than expected ($\chi^2 = 29.01$, df = 5, $p < .001$), while the Dynata and student samples contained more straight-lining than expected. Based on a one-way ANOVA test using a Tukey’s post hoc analysis, the personal reliability of the MT0 sample was lower than MTM, Dynata, and student samples (f (5, 1294) = 5.26, $p < .001$, with mean differences significant at $p < .05$ or better). Finally, the format of the survey (putting questions together versus separating them by construct) influenced responses for the MTM, Qualtrics and student samples.

Table 6
Statistical accuracy screener findings.

	MT0[a]		MTL[a]		MTM[a]		Qualtrics		Dynata		Students	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Total	200	100%	201	100%	154	100%	212	100%	418	100%	134	100%
Fast	18	9%	19	10%	22	11%	4	2%	17	9%	5	3%
Slow	4	2%	4	2%	2	1%	3	2%	5	3%	4	2%
SL	22	11%	21	10%	22	14%	32	15%	101	24%	28	21%
PR	0.68		0.77		0.86		0.78		0.81		0.78	
Format	ns		ns		0.10*		0.12*		ns		0.20*	

Fast indicates number and percentage of respondents who spent less than 80% of the estimated 13 min on the survey (<156 s).
Slow indicates number and percentage of respondents who spent more than 80% of the estimated 13 min on the survey (>1404 s).
SL = Straight-lining, representing the number and percentage of respondents who rated all items identically (e.g., "4″ for every item).
PR = Personal reliability, representing the Spearman's Rho correlation between split halves of the perceived similarity scale.
Format = Difference in reliability between the mixed and separated survey formats.
* Denotes a statistical difference at p < .05.
[a] MTurk samples: MT0 = no accuracy screeners, MTL = low accuracy screeners, MTM = high accuracy screeners

In sum, the results were quite inconsistent. While the MTurk samples were more likely to speed, they were less likely to engage in straight-lining. The personal reliability of the MT0 sample was lower than many of the other samples, but survey format did not influence their responses.

5.5. Redundancy between screeners

This section discusses redundancy between screener types. Table 7 shows correlations among the direct screeners. High magnitude in correlation suggests an overlap among screeners. Overlap can also be measured at the individual level. Table 8 shows a count of the number of screeners that each respondent failed. Dummy variables were created for each screener (0 = passed, 1 = does not pass). If respondents fail two screeners, then they would have a "1 = does not pass" for each dummy variable.

As shown in Table 8, direct-selection screeners do not necessarily remove the same participants as direct-accuracy screeners. A total of 83% of MT0 respondents failed at least one researcher-generated, direct-selection screener, 61% failed at least one researcher-generated, direct-accuracy screener, and 91% failed at least one direct screener of either type. Consequently, the selection and accuracy screeners are not redundant. Within each category, the selection screeners display apparent redundancy as shown by the greater number of respondents who failed more than one screener. Thus, picking a single effective selection screener should be sufficient. However, for accuracy screeners, the majority of respondents failed on a single screener. Consequently, the choice of accuracy screener affects the set of respondents included in the sample.

The percentage of respondents screened is an indicator that the sample contains a significant percentage of insufficient quality responses. In other words, few respondents were able to meet the standards of inclusion. However, the remaining sample should be considered of adequate quality, regardless of how much of the sample was removed. Thus, while only a small percentage of respondents from MT0 remained after all of the screeners were applied while a much higher quantity of Qualtrics respondents remained, those remaining MT0 respondents should be considered just as acceptable quality for further analysis as the Qualtrics sample. What may be necessary for researchers is to set their expected sampling quotas higher knowing that there will be lost datapoints in order to achieve a large enough sample for proper analysis.

5.6. Responses uniqueness and contamination

Previous research has found that MTurk respondents participate in studies repeatedly if studies are posted multiple times (Chandler et al., 2014). However, research has not investigated other forms of cross-contamination. In this research, only two instances were found of respondents taking the survey multiple times using different platforms (MTurk and Qualtrics, for example). Within the MTurk platform, about 18% of MTurk respondents participated in both the MT0 and MTL samples. Within the same sample, the MTM sample showed 18.2% percent duplicate IP-addresses, though this could be explained by multiple people in the same household or location having master-level accounts. Similarly, the student data showed a substantial number of duplicate IP-addresses, which is likely explained by student use of shared computer labs. However, no student participated in a different sample. There was only a single comment in the study purpose that indicated a respondent saw both versions of the "red-herring" experimental design in which respondents were shown either a white or black service provider.

However, an online search showed that the present study had been discussed on user forums. Because MTurkers from the non-Master samples were paid fifty cents, and this posting indicates that the pay is $1.50, it must have been posted by a Master MTurker. It did not appear that the purpose of the study was discussed. However, respondents commented on other studies. For example, this comment was posted on a study about the similarity-attraction effect: "good ol' fashioned 'y u so racist and/or sexist?' Loaded choices." Forum comments like these have the potential to bias results not only for that particular study but for all studies listed on the researcher's profile.

Table 7
Correlation among direct screeners.

	Screener	1	2	3	4	5	6	7	8	9
1	IP Outside US									
2	Location mismatch	0.278**								
3	Self-report Southern state	−0.126**	−0.271**							
4	Zip code match	0.084**	0.262**	−0.211**						
5	State capital incorrect	0.146**	0.388**	−0.286**	0.410**					
6	Picked 2	−0.036	−0.101**	−0.027	−0.069*	−0.169**				
7	Oak Tree Pass	−0.002	−0.095**	0.024	−0.145**	−0.203**	0.224**			
8	Honesty test	0.004	−0.182**	0.140**	−0.144**	−0.194**	0.002	0.034		
9	Not an income	−0.023	0.072**	0.002	0.076**	0.137**	−0.093**	−0.058*	−0.006	
10	Not a race	0.035	0.165**	−0.084**	0.163**	0.203**	−0.113**	−0.091**	0.020	0.377**

Table 8
Redundancy between screeners.

		Failed Accuracy						Failed Selection						Failed Total					
Sample	*n*	0		1		2+		0		1		2+		0		1		2+	
MT0	200	78	39%	96	48%	26	13%	34	17%	51	26%	115	58%	19	10%	45	23%	136	68%
MTL	201	102	51%	77	38%	22	11%	42	21%	56	28%	103	51%	31	15%	43	21%	127	63%
MTM	154	118	77%	36	26%	0	0%	85	55%	53	38%	16	12%	67	44%	60	44%	27	20%
Qualtrics	212	95	45%	87	42%	30	14%	141	67%	54	26%	17	8%	69	33%	79	38%	64	31%
Dynata	418	184	44%	167	40%	67	16%	288	69%	97	23%	33	8%	141	34%	150	36%	127	30%
Students	134	55	41%	54	41%	25	19%	113	84%	13	10%	8	6%	50	37%	47	35%	37	28%

6. Recommendations

This research developed a typology of screeners based on sampling goal and methodology and then applied the typology to samples from several online panels. The typology looked at the sampling goals of *selection*, which is whether the sample is from the target population, and *accuracy*, which is whether respondents are attentive. *Direct screeners* evaluate individual respondents while *statistical screeners* evaluate the overall sample. By providing a framework of screeners, a key contribution of this research is to provide guidance for the design of new screeners.

6.1. Using and reporting screeners

Researchers should use one or two researcher-generated screeners for each screening category: (1) direct selection, (2) statistical selection, (3) direct accuracy, and (4) statistical accuracy. Our findings indicate that selection and accuracy screeners are not redundant; hence, researchers must use both. Each category of screener has important implications for achieving sampling goals. Selection screeners ensure that the sample matches the target population while accuracy screeners ensure honest and attentive responses. Direct screeners have the advantage of immediately identifying problematic responses without the need for further analysis but the disadvantage of being visible to participants. Conversely, statistical screeners are unobtrusive to participants (DeSimone et al., 2015) but require further analysis. No single category of screener can adequately screen data so researchers should use them in combination.

For selection screeners, respondents should be removed when they are not actually a member of the target population, and for accuracy screeners, the respondents should be removed when they are not honest and attentive. The question is whether researchers can know the "truth" about selection and accuracy to ensure that inappropriate responses are removed without inadvertently removing appropriate responses. Because direct screeners have clear "right" and "wrong" answers, it is very unlikely that direct screeners will falsely lead to the removal of appropriate responses. Respondents failing direct screeners are either not a member of the target population or are insufficiently honest/attentive and should be removed. However, because direct screeners are visible to respondents, it is easier for them to circumvent direct screeners than statistical screeners. As a result, direct screeners may not identify all inappropriate responses.

Statistical screeners have the advantage of being largely invisible to respondents. However, they also require additional consideration. Ideally, the person responsible for applying statistical screeners (e.g., the project manager or researcher) should be separate from the researcher who analyzes the theoretical relationships. Furthermore, the thresholds for excluding data should be based on an objective standard whenever possible. Researchers might compare the sample to the known characteristics of a target population to determine match or apply objective criteria, such as a synonym-antonym analysis or a comparison of response time on a task that has a known duration. When an objective standard is not readily available, statistical screeners can still be used to provide a ''second line of defense'' for assessing sample quality but should only be used to identify and remove individual responses when there is a clear and defensible justification.

The supplier-provided screening criteria provided by Amazon, including number of HITs completed, percentage of HITs approved, and Master designation, did somewhat overlap with the research-generated direct accuracy screeners used in this research. However, the researcher-generated screeners flagged many additional respondents for removal above and beyond the supplier-provided screeners, suggesting that researchers should include at least some screeners for each category when collecting data from online panels. Unfortunately, most studies listed in Table 1 do not discuss how respondents were screened. Researchers should screen their data for quality and provide a full account of such procedures.

The most effective screeners depend on the research materials. For instance, excessive response time may indicate an interruption, which would not inherently lead to a poor-quality response. However, tasks interruptions can disrupt reading comprehension (Foroughi, Barragan, & Boehm-Davis, 2015), which might influence how respondents react to experimental stimuli. Hence, there can be justification for removing responses based on excessive response time depending on the research materials.

The data cleaning process can affect correlations among variables, but it is unclear whether they would increase or decrease correlations in a particular data set. Noisy data can attenuate valid correlations, but poor-quality data may also cause systematic correlational biases (e.g., the halo error) and/or acquiescence bias which may lead to ''illusionary'' correlations between variables (Podsakoff, MacKenzie, & Podsakoff, 2012). In order to assess the impact of data cleaning on correlations, sample results could be compared to known population values, instead of relying on internal sample correlations. Here, we could compare sample reports of motorcycle ownership to population estimates taken from motor vehicle registrations. Similarly, a survey could contain questions about home ownership, age of oldest child, number of children, or any other variables for which population correlation values could be computed from census data. Sample results could be compared with matching values supporting data quality.

Furthermore, the effectiveness of any particular screener is likely to change over time. This is particularly true for ''professional survey-takers'' like MTurk masters because they maintain discussion boards. In our research, the commonly used ''pick 2'' attention check removed the fewest respondents from the master MTurk sample, followed by the other MTurk samples, ostensibly because MTurk respondents had encountered that type of attention check previously. To ensure that screeners are effective for experienced survey-takers, researchers may need to continually update the details of their screening questions to stay ahead of the respondents. Thus, rather than relying on a few commonly used screeners, researchers must adapt their screeners in a sort of ''arms-race.'' This ''arms-race'' will likely accelerate as both respondents and researchers use more sophisticated artificial intelligence. A key benefit of the typology presented in this research is to provide a framework for researchers to use for continually developing new and better screeners for each category of screener (selection/accuracy and direct/statistical).

6.2. Ethical use of screeners

Researchers should report the use and results of all four types of screeners. Researchers should report how many respondents were removed per screener. Direct screeners should be used to remove poor quality responses prior to hypothesis testing or other topical analysis; thus, researchers should be unaware of how direct screeners might have influenced results. After direct screeners have been applied, statistical screeners can provide insight into the achievement of sampling goals. Sampling quality screeners should be selected prior to data collection and be the sole determination of how well sampling goals were achieved: Theoretical study results should never be used to determine whether a sample provides sufficient quality. If LQD (low quality data) remain, sampling quality screeners can be used to remove the offending data, but this should be done prior to hypothesis testing and all other analyses addressing research questions. Ideally, the primary data analyst who screens the data should be separate from the primary analyst who tests hypotheses. Data transparency is a best practice and coauthors should replicate each other's analyses.

As mentioned previously, the most appropriate outlier analysis for a data set depends on the characteristics of the data (Aggarwal, 2017; Woolrich, 2008). For example, standard scores should only be used to identify outliers when the sample data is normally distributed and is known to approximate the distribution of the target population. Thus, choice of outlier analysis should be justified.

6.3. Recommended crowdsourcing services

The actively managed crowdsourcing services, Dynata and Qualtrics, had higher selection sampling quality than the MTurk samples, defined in this research as matching the target population of adults living in the Southern U.S. The unscreened MTurk samples had very low selection quality, and most of the MTurk sample failed one or more direct selection screeners. While the master sample performed the best of the three samples, fewer than 50% of respondents appear to be representative of the target population compared to 66% of the Qualtrics, 69% of the Dynata, and 84% of the respondents in the student samples. For accuracy sampling quality, the actively-managed crowdsourcing services were less likely to speed than all of the MTurk samples, more likely to pass direct accuracy screeners than the non-master MTurk samples, and had higher personal reliability than the non-master MTurk samples. However, they engaged in high levels of straight lining, particularly the Dynata sample, in which 24% of the unscreened sample selected a single response for every item. That said, actively-managed panels will typically do some screening beforehand so that researchers will not receive these poor-quality responses. Our research sought to evaluate all completes, but researchers can expect "good completes" to be screened. Preferably, the data service provider does most of the screening because they are not involved in publication.

The high percentage of MTurk respondents removed by statistical selection screeners was not unexpected. A key weakness of all minimally-managed online samples is that respondents are able to self-select into studies. MTurk respondents also had a relative high number of duplicate IP-addresses, second only to the student sample in which respondents were using a common computer lab. Another weakness of self-selection is that it allows respondents to participate in multiple, repeated surveys. If researchers use MTurk samples for multiple, related studies, many respondents will be duplicates. For experimental designs involving manipulated conditions, purportedly between-subjects experiments take on the demand characteristics of within-subjects designs using repeat respondents. Even for surveys, repeated exposure to the same items can cause testing bias (Shadish, Cook, & Campbell, 2002).

As a result of these weaknesses, MTurk samples require rigorous direct and statistical selection screeners to ensure that participants are from the target population. Only 17% of MT0 respondents successfully passed all of the direct-selection screeners; thus, researchers using a similar sample for a similar population would need to request more than five times as many respondents to achieve a desired sample size. As a result, the cost may be comparable to purchasing from an actively-managed service, yet, there may be more problems that go undetected. Because of these drawbacks, it is recommended that MTurk samples be avoided when selection sampling quality is a primary concern (e.g., targeting professional salespeople, a particular demographic, people with a specific hobby, etc.), or when materials are replicated across multiple studies.

For statistical accuracy screeners, the results were mixed. As suggested by Ford (2017), the MTurkers were more prone to speeding than the other samples. The non-master samples were also more prone to failing direct accuracy screeners and had lower personal reliability. However, those samples were also the least likely to engage in straight-lining behavior and less susceptible to survey format, though given that they were more likely to speed through the survey than other samples, this might suggest the use of harder to identify response biases, such as Christmas trees or arbitrary responses. Master MTurkers were less likely to report inconsistencies about their identity and successfully pass direct accuracy screeners. Thus, Master MTurk samples appear to be more attentive and more consistent than the other two MTurk samples.

6.4. Targeting a narrow population

A key concern is the use of panel data when sampling a specific subgroup of people (e.g., salespeople). Many panels will not contain enough sampling units matching specific subgroups for large studies or repeated studies using different respondents. A secondary problem related to this issue is that researchers in certain disciplines are likely to overrepresent the small body of respondents from these panels who claim to be members of the target subgroup. Our research suggests that researchers should use multiple panels rather than relying on a single panel. From our findings, it does not appear that respondents are prone to using multiple platforms. Therefore, rather than drawing on a single platform repeatedly, switching among platforms should reduce overreliance on a few subgroup "superworkers." Additionally, researchers should try to use broader subgroups whenever possible to access a larger number of respondents. Nevertheless, our findings also indicate that many respondents from minimally-managed panels are willing to mislead researchers about their identity; hence, it is important to use selection screeners in combination with broadening the target subgroup.

7. Limitations and directions for future research

The results reported in this research provide insight into how many respondents would be screened from each sample for each category of screener. While it is likely that researchers using similar screening criteria would find similar results for each sample, respondents belonging to the same source are not necessarily homogeneous. The composition of the sample pool may change over time, and the effectiveness of screening criteria may change. Thus, the generalizability of the specific findings of this research may be somewhat limited, although the broader implication of how researchers can design, use, and report each category of screener should still apply. Future research should also explore other actively-managed and minimally-managed crowdsourcing services.

While the screener typology presented in this research can and should be used for all human subject sampling frames, the specific screening questions will need to be adapted to fit the needs of the researcher. Additionally, evidence suggests some weaknesses for all online samples. More research needs to be conducted to explore how new technologies have influenced survey response. The Qualtrics software provides proprietary algorithms for estimating the likelihood of fraud using the captcha verification question and respondent meta-data (i.e., Relevant ID). We did not explore the use of these tools because they are only available on certain Qualtrics software licenses. Hence, they are not widely available. Yet, future research should estimate the effectiveness of these algorithms compared to other methods of screening data, particularly given our finding that the captcha verification question did not eliminate any bots. Furthermore, actively-managed online data collection services typically prescreen data but the data from Dynata and Qualtrics presented in this research were not prescreened. Researchers following normal policies should expect to receive higher-quality "good-completes" from these sources.

Pertaining to Amazon's Mechanical Turk, Master MTurkers were better able to successfully pass direct accuracy screeners than the other two MTurk samples; yet evidence of their statistical accuracy remains mixed. Ability to pass accuracy direct screeners may reflect an over-sensitization bias caused by taking so many surveys or a concerted effort to "play the game." Many online crowdsourcing services prohibit direct contact between respondents and researchers through specific communication mediums, such as by phone or email. For example, Prolific asks respondents to immediately report if a researcher asks for personal contact information. Nevertheless, it would be worthwhile to know more about the respondents "behind the scenes." Finally, more work is needed to learn why the non-master MTurk samples were less likely to engage in straight-lining and less susceptible to survey format. Perhaps more sophisticated methods of detecting other response bias patterns can be developed and used as an additional accuracy statistical screener.

8. Conclusion

The present research advises the use of screeners for selection to help ensure that samples are drawn from the targeted population, and for accuracy, which assesses the extent to which respondents provide honest, attentive responses. Furthermore, we used two screener methodologies, direct screeners that remove respondents based on criteria embedded into the research materials and statistical screeners that use analyses invisible to respondents. Data were screened using each of the four screener types: direct selection, statistical selection, direct accuracy, and statistical accuracy. Using these criteria, at least some of the responses were screened from every sample. Researchers should always report all screening procedures, including all selection and accuracy screeners, and they need to be specific about when the screening actually took place. Response rate becomes lower as more data has to be removed from that initially collected to obtain a valid sample. In that way, we should return to the practice of accounting for potential non-response bias (Armstrong & Overton, 1977). Furthermore, reviewers should always assess screener quality and avoid "punishing" researchers for accurately reporting how they screened for poor quality data. The responsible use of screening criteria will improve online data quality, and ultimately, the replicability of marketing findings.

References

Aggarwal, C. C. (2017). *Outlier analysis in Data Mining* (pp. 237–263). Cham: Springer.

Armstrong, J. S., & Overton, T. S. (1977). Estimating nonresponse bias in mail surveys. *Journal of Marketing Research, 14*(3), 396–402.

Babin, B. J., Griffin, M., & Hair, J. F. Jr., (2016). Heresies and sacred cows in scholarly marketing publications. *Journal of Business Research, 69*(8), 3133–3138.

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.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaivete among Amazon Mechanical Turk Workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130.

Chesney, T., & Penny, K. (2013). The impact of repeated lying on survey results. *SAGE Open, 3*(1).

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.

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

Conte, A., Levati, M. V., & Montinari, N. (2014). Experience in public goods experiments. No. 2014-010. Economic Research Papers.

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). MTurk workers' use of low-cost virtual private servers to circumvent screening methods: A research note.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*(2), 171–181.

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 Psychology, 33*, 559–577.

Downs, J. S., Holbrook, M. S., & Peel, E. (2012). *Screening participants on Mechanical Turk: Techniques and justifications*. ACR North American Advances.

Feick, L., & Higie, R. A. (1992). The effects of preference heterogeneity and source characteristics on ad processing and judgements about endorsers. *Journal of Advertising, 21*(2), 9–24.

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

Foroughi, C. K., Werner, N. E., Barragán, D., & Boehm-Davis, D. A. (2015). Interruptions disrupt reading comprehension. *Journal of Experimental Psychology: General, 144*(3), 704–709.

Fricker, R. D. (2008). Sampling methods for web and e-mail surveys. The SAGE handbook of online research methods (pp. 195–216).

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

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis*. Andover, U.K: Cengage.

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.

Hauser, D., Paolacci, G. & Chandler, J. J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In F. Kardes, P. Herr, & N. Schwarz (Eds.), Handbook in research methods in consumer psychology. New York/London: Routledge.

Huff, C., & Tingley, D. (2015). Who are these people? Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics, 2*(3), 1–12.

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*, 99–114. https://doi.org/10.1007/s10869-011-9231-8.

Hulland, J., & Miller, J. (2018). Keep on Turkin? *Journal of the Academy of Marketing Science, 46*(5), 789–794.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141–155.

Matherly, T. (2019). A panel for lemons? Positivity bias, reputation systems and data quality on MTurk. *European Journal of Marketing, 53*(2), 195–223.

Montoya, R. M., Horton, R. S., & Kirchner, J. (2008). Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships, 25*(6), 889–922.

NationMaster (2019). Philippines vs United States cost of living stats compared. Retrieved from https://www.nationmaster.com/country-info/compare/Philippines/United-States/Cost-of-living.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411–419.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569.

Rand, D. G. (2012). The promise of Amazon Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172–179.

Reddit.com: Short consumer survey (~10 minutes) - Broken attention check https://www.reddit.com/r/mturk/comments/75prll/short_consumer_survey_10_minutes_broken_attention/. Last accessed 10-31-19.

Robinson, J., Rosenzweig, C., Moss, A. J., & Litman, L. (2019, June 7). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. https://doi.org/10.1371/journal.pone.0226394.

Shadish, C., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science, 1*(2), 213–220.

Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research, 69*(8), 3139–3148.

Wessling, K. Sharpe, Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research, 44*, 211–230.

Woolley, K., & Risen, J. L. (2021). Hiding from the Truth: When and how cover enables information avoidance. *Journal of Consumer Research, 47*(5), 675–697.

Woo, S. E., Keith, M., & Thornton, M. A. (2015). Amazon Mechanical Turk for industrial and organizational psychology: Advantages, challenges, and practical recommendations. *Industrial and Organizational Psychology, 8*(2), 171–179.

Wood, D., Harms, P., 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.

Wood, J. A., Boles, J. S., Johnston, W., & Bellenger, D. (2008). Buyers' trust of the salesperson: An item-level meta-analysis. *Journal of Personal Selling & Sales Management, 28*(3), 263–283.

Woolrich, M. (2008). Robust group analysis using outlier inference. *Neuroimage, 41*(2), 286–301.