Journal of Management
Vol. 47 No. 4, April 2021 823–837
DOI: 10.1177/0149206320969787

MTurk Research: Review and Recommendations

Herman Aguinis
Isabel Villamor
The George Washington University
Ravi S. Ramani
Morgan State University

The use of Amazon's Mechanical Turk (MTurk) in management research has increased over 2,117% in recent years, from 6 papers in 2012 to 133 in 2019. Among scholars, though, there is a mixture of excitement about the practical and logistical benefits of using MTurk and skepticism about the validity of the data. Given that the practice is rapidly increasing but scholarly opinions diverge, the Journal of Management *commissioned this review and consideration of best practices. We hope the recommendations provided here will serve as a catalyst for more robust, reproducible, and trustworthy MTurk-based research in management and related fields.*

Keywords: *Amazon Mechanical Turk; research methods; online data collection; research design; MTurk; experiments*

The use of web-based research using Amazon's Mechanical Turk (MTurk) has increased tenfold over just the last decade (Walter, Seibert, Goering, & O'Boyle, 2019), making it by far the most frequently used online data collection method (Porter, Outlaw, Gale, & Cho, 2019). Despite its popularity, there are concerns that call into question the validity of research conclusions based on MTurk data (e.g., Barends & de Vries, 2019; Hydock, 2018; Zack, Kennedy, & Long, 2019). These concerns are severe enough that some journals have

---

Portions of this article were presented at the meeting of the Academy of Management, Boston, MA, August 2019.

Supplemental material for this article is available with the manuscript on the JOM *website.*

Corresponding author: Herman Aguinis, Department of Management, School of Business, The George Washington University, Funger Hall, Suite 311, 2201 G St. NW, Washington, DC 20052, USA.

E-mail: haguinis@gwu.edu

intermittently refused to accept manuscripts that utilized MTurk, and some journal editors and reviewers have summarily recommended rejecting manuscripts that used MTurk regardless of a study's other positive features (Landers & Behrend, 2015; Walter et al., 2019). Given the growth in the use of MTurk for research in management and related fields, and persistent concerns about its trustworthiness, the *Journal of Management* tasked our team with reviewing the MTurk-based literature and developing actionable best-practice recommendations for using this data collection tool.

Literature Review

Scoping Substantive Review

We began with a scoping substantive review (Paré, Trudel, Jaana, & Kitsiou, 2015) to capture the breadth of the literature regarding MTurk use. We included empirical papers drawn from 15 journals that used MTurk to collect data, test hypotheses, or validate scales. Details about our review's scope as well as journal and article selection procedures are included in online supplement Appendix A. Between January 2005 and May 2020, 510 empirical papers using MTurk samples have been published in the journals we examined (see online supplement Appendix B for details). Moreover, the use of MTurk has grown markedly ($R^2 = .94$ for the linear trend; see online supplement Appendix C for details). From 6 articles in 2012 to 133 in 2019, the use of MTurk has increased over 2,117%.

Critical Methodological Review

We followed with a systematic and transparent five-step process to identify methodological sources about MTurk based on existing best-practice recommendations (see online supplement Appendix D for details; Aguinis, Ramani, & Alabduljader, 2018, in press). We began with 32 journals, and the final list upon which we base our best-practice recommendations includes 144 sources (119 articles published in 65 journals, 23 presentations from 11 conferences, a working paper, and a book). Online supplement Appendix E lists these sources, the number of items drawn from each, and the individual items. In the interest of transparency and replicability, online supplement Appendix F lists the 96 items that were initially considered but eventually excluded.

Summary of Findings

Based on these literature reviews, we determined that MTurk's popularity can be broadly attributed to four closely related benefits compared with research conducted using more traditional samples: (a) large and diverse participant pool, (b) ease of access and speed of data collection, (c) reasonable cost, and (d) flexibility regarding research design choice. We describe each of these benefits in Table 1. We also determined that there is justifiable skepticism due to unique challenges that pose validity threats to substantive conclusions. Specifically, we identified 10 particularly salient challenges of MTurk research: (a) inattention, (b) self-misrepresentation, (c) self-selection bias, (d) high attrition, (e) inconsistent English language fluency, (f) non-naivete, (g) growth of MTurker communities, (h) vulnerability to web robots (or "bots"), (i) social desirability bias, and (j) perceived researcher unfairness. Some of these challenges also apply to other data collection methods (e.g.,

Table 1
Summary of Main Benefits of Using Amazon Mechanical Turk (MTurk) for Conducting Management Research

Benefit	Description of Benefit
1. Large and diverse participant pool[3,4,5,9,12,15,20]	1. MTurk allows researchers access to a larger and more demographically diverse participant pool as compared with traditional student samples and the U.S. population. Compared with traditional student samples, MTurkers are older, have more years of relevant work experience, and report greater computer and internet knowledge. Compared with the general U.S. population, MTurkers are younger and more educated. In addition, demographic and political-affiliation differences can be eliminated by controlling for 10 factors (i.e., age, gender, race, ethnicity, income, education, marital status, religion, ideology, and political partisanship). Thus, MTurk has the potential to complement laboratory studies by ensuring the transportability of results.
2. Ease of access and speed of data collection[6,7,11,13,16]	2. About 7,300 MTurkers are available for a study at any given time. By maintaining a relatively stable large online pool of participants, MTurk greatly reduces recruitment efforts, thereby making it easier to conduct, extend, reproduce, replicate, or modify a study. Most MTurk assignments are completed within 12 hours or less.
3. Reasonable cost[6,10,11,13,14]	3. Researchers can gather data at a lower cost than when using samples of students or working adults or using participants recruited through other online panel websites. MTurk's constant fee structure (i.e., the amount paid to Amazon to conduct a study) and integrated payment infrastructure reduces considerably the administrative costs associated with compensating participants.
4. Flexibility regarding research design choice[1,2,6,8,13,14,17,18,19]	4. MTurk can be used to implement experimental, passive observation, quasiexperimental, and longitudinal research designs and even perform tasks such as content analysis. Furthermore, MTurk can be used to conduct cross-cultural and international research by restricting the participant pool to workers with specific cultural backgrounds or to those who live in particular countries. Together, these benefits allow researchers to advance theory by testing hypotheses in diverse samples and about different types of effects and relations between variables (e.g., upward and downward, over time, dyadic).

Note: Sources used to summarize benefits: [1]Alonso and Mizzaro (2012); [2]Arechar, Gächter, and Molleman (2018); [3]Bader, Baumeister, Berger, and Keuschnigg (2020); [4]Behrend, Sharek, Meade, and Wiebe (2011); [5]Berinsky, Huber, and Lenz (2012); [6]Buhrmester, Talaifar, and Gosling (2018); [7]Bunge et al. (2018); [8]Callison-Burch and Dredze (2010); [9]Casler, Bickel, and Hackett (2013); [10]Chandler, Rosenzweig, Moss, Robinson, and Litman (2019); [11]Heer and Bostock (2010); [12]Levay, Freese, and Druckman (2016); [13]Mason and Suri (2012); [14]Paolacci, Chandler, and Ipeirotis (2010); [15]Pearl and Bareinboim (2014); [16]Stewart et al. (2015); [17]Stritch, Pedersen, and Taggart (2017); [18]Summerville and Chartier (2013); [19]Tosti-Kharas and Conley (2016); [20]Weinberg, Freese, and McElhattan (2014).

laboratory studies relying on students, field studies sampling working adults), but the validity threats they pose are even more salient when using MTurk. Table 2 describes these challenges of MTurk research and associated validity threats.

Recommendations

In view of our findings, we provide 10 best-practice recommendations organized around the three typical stages of an empirical study: planning, implementation, and reporting of results. Table 3 summarizes each of the recommendations and the particular MTurk challenge(s) addressed by each. While some of these best practices may also apply to non-MTurk studies, our checklist focuses specifically on how to mitigate validity threats when using MTurk.

Table 2
Challenges of Amazon Mechanical Turk (MTurk) Research and Associated Validity Threats

Challenge	Description	Associated Validity Threat(s)
1. MTurker Inattention[3,8,9,12,13,18,21]	1. MTurkers often complete HITs in distracting environments and at rapid speed to maximize monetary returns, which translates into about 15% of MTurkers failing attention and compliance checks. MTurkers are less likely to pay attention to study instructions or manipulations, and more likely to engage in insufficient effort or careless responding, as compared with college student samples. Compared with student samples, online participants are significantly more likely to be distracted due to cell phone use (MTurker = 21% vs. student = 9%), internet surfing (MTurker = 11% vs. student = 1%), or conversing with another person (MTurker = 21% vs. student = 2%).	• Internal validity • Construct validity • Statistical conclusion validity
2. Self-misrepresentation[9,19,20,23,24]	2. MTurkers may misrepresent self-reported demographic, personality, and other characteristics to meet a study's eligibility criteria. Estimates of the percentage of MTurkers who engage in such practices range from 10% to 13%, to 24% to 83%. The most commonly misrepresented characteristics are income (38.2%), education (31.3%), age (22.6%), family status (14.8%), and gender (6.6%).	• External validity
3. Self-selection bias[12,13]	3. Unlike traditional samples, where the researcher defines the potential participant pool (e.g., first-line managers at a company), the decision to be an MTurker is based on an individual's personal and demographic characteristics, such as monetary incentives, boredom, employment status, or country location. These characteristics, which can serve as confounds and alternative explanations for observed relations, compromise the researchers' ability to randomly sample from their target population and therefore pose a threat to external validity.	• External validity
4. High attrition rates[2,9,12,25]	4. Attrition rates in MTurk studies often exceed 30% (range: 31.9%–51%). The online nature of MTurk studies leads to higher attrition rates than laboratory experiments or field research and even the possibility of differential attrition.	• Internal validity • External validity
5. Inconsistent English language fluency[15,18]	5. English language fluency influences how participants interpret the study's instructions, manipulations, and measures. Data from MTurkers from countries where English is not the primary language displays only configural invariance with data collected from undergraduates and organizational employees from countries where English is the primary language.	• Internal validity • Construct validity • Statistical conclusion validity
6. MTurker non-naivete[9,10,11,12]	6. While MTurk's software prevents participants from receiving compensation more than once for the same study, it does not track participant exposure to studies that examine particular topics or, even worse, use the exact same stimuli or manipulation. A small number of MTurkers (10%) account for over 40% of completed studies, and many participants "specialize" in studies that examine specific topics or are conducted by the same researchers. Accordingly, many MTurkers are familiar with experimental settings and tasks (e.g., framing alternatives for decision-making scenarios, using videos to manipulate emotions) and research materials (e.g., measures, vignettes), which can, on average, reduce effect size estimates by up to 40%.	• Internal validity • Construct validity

(continued)

Table 2 (continued)

Challenge	Description	Associated Validity Threat(s)
7. Growth of MTurker communities[7,10,12]	7. 61% of MTurkers interact with other participants regarding their experience. Thus, MTurkers are often aware of a study's purpose or the manipulations used.	• Internal validity • Construct validity
8. Vulnerability to web robots (or "bots")[8]	8. Web robots (or "bots") are malicious software programs designed to specifically participate in online studies to receive compensation. These programs, which are often freely available and easy to use, generate data that follow a random or partially random distribution in response to a study's questions, thereby making it harder to distinguish between web robots and inattentive or careless participants. While we currently lack estimates of the percentage of MTurk data attributable to web robots, such programs represent a feature that can impact research conducted using MTurk.	• Internal validity • Construct validity • Statistical conclusion validity
9. MTurker social desirability bias[1,5,12,22]	9. Because monetary compensation is one of the primary reasons for participating in a HIT, MTurkers are more likely to provide socially desirable responses than student samples. The percentage of respondents who engage in this practice varies across countries, with U.S. participants more likely to provide socially desirable responses compared with Indian participants.	• Internal validity • Construct validity
10. Perceived researcher unfairness[4,6,7,9,12,14,16,17]	10. In addition to concerns about the fairness of procedures used to make compensation decisions, issues that cause MTurkers to perceive researchers as unfair include a lack of a process to communicate with researchers, unavailability of disability access features, and inaccurately stated time requirements. Participants who feel treated unfairly can share their experiences in MTurker communities, leading to punitive actions, such as a boycott of subsequent studies by that researcher.	• External validity

Note. Sources used to derive recommendations: [1]Antin and Shaw (2012); [2]Arechar, Gächter, and Molleman (2018); [3]Barends and de Vries (2019); [4]Bederson and Quinn (2011); [5]Behrend, Sharek, Meade, and Wiebe (2011); [6]Bergvall-Kåreborn and Howcroft (2014); [7]Brawley and Pury (2016); [8]Buchanan and Scofield (2018); [9]Buhrmester, Talaifar, and Gosling (2018); [10]Chandler, Mueller, and Paolacci (2014); [11]Chandler, Paolacci, Peer, Mueller, and Ratliff (2015); [12]Cheung, Burns, Sinclair, and Sliter (2017); [13]Clifford and Jerit (2014); [14]Deng, Joshi, and Galliers (2016); [15]Feitosa, Joseph, and Newman (2015); [16]Fieseler, Bucher, and Hoffmann (2017); [17]Gleibs (2017); [18]Goodman, Cryder, and Cheema (2013); [19]Hydock (2018); [20]Kan and Drummey (2018); [21]Litman, Robinson, and Rosenzweig (2015); [22]Mummolo and Peterson (2019); [23]Necka, Cacioppo, Norman, and Cacioppo (2016); [24]Wessling, Huber, and Netzer (2017); [25]Zhou and Fishbach (2016). HIT = human intelligence task.

Planning Stage

1. Evaluate appropriateness of MTurk to develop or test theories. Our first recommendation is to evaluate the alignment between the desired target population and that of MTurkers and collect and report detailed sample characteristics rather than assume similarity with earlier MTurk studies (Chandler & Paolacci, 2017). This helps address challenges associated with self-selection bias (Casey, Chandler, Levine, Proctor, & Strolovitch, 2017). For example, MTurkers show differences compared with laboratory samples on Big Five personality traits (Colman, Vineyard, & Letzring, 2018). Therefore, when Big Five traits are expected to influence substantive results, they can be used as statistical controls so that results and inferences are attributable to the hypothesized predictors and not to variability in personality traits between samples (Bernerth & Aguinis, 2016).

Table 3

Summary of Best-Practice Recommendations for Addressing Validity Threats in Research Using Amazon Mechanical Turk (MTurk)

Stage of Study	Recommendation	Implementation Guidelines	MTurk Challenge(s) Addressed (From Table 2)
Planning	1. Evaluate appropriateness of MTurk to develop or test theories	✓ Evaluating alignment between desired target population and that of MTurkers ✓ Collecting and reporting detailed sample characteristics rather than assuming similarity with earlier MTurk studies	• Self-selection bias
	2. Decide qualifications used to screen MTurkers	✓ Deciding qualifications (e.g., age, work experience, race) relevant to study ✓ Evaluating MTurkers using a screener study, paying everyone who participates, eliminating those who do not match the desired criteria, and inviting those who meet the qualifications/pass the screener to participate in the focal study ✓ Determining a priori whether to consider only MTurkers from native-English-speaking countries (based on their internet protocol [IP] addresses) or to establish measurement equivalence across native and non-native English speakers ✓ Deciding whether to use only highly qualified MTurkers (i.e., "Master Workers") or to employ screening questions to gauge MTurker familiarity with research subject, stimuli, and, if applicable, manipulations	• Self-misrepresentation • Inconsistent English language fluency • MTurker non-naivete
	3. Establish required sample size	✓ Planning to collect data from at least an additional 15% to 30% of MTurkers to compensate for participant attrition and failure to pass attention checks	• MTurker inattention
	4. Formulate compensation rules	✓ Paying U.S. minimum wage when drawing on U.S. samples ✓ Deciding a priori what criteria (if any) will be used to refuse payment to MTurkers ✓ Using a consent form, including details on compensation rules (i.e., codes of conduct, monitoring procedures, and penalties for fraudulent or untruthful reporting; see online supplement Appendix G for a customizable template)	• High attrition rates • Perceived researcher unfairness
	5. Design data collection tool used to gather responses	✓ Requiring MTurkers to complete an informed consent form, including a "CAPTCHA" verification to thwart web robots (or "bots") ✓ Requiring MTurkers to provide their MTurk ID and maintaining a reference database of past participants to identify MTurkers who attempt self-misrepresentation ✓ Using at least two attention checks (e.g., instructed items that direct MTurkers to complete or abstain from a particular action, bogus items that ask MTurkers to answer obvious or ridiculous questions, self-reports of effort, and questions on which all or almost all respondents should provide the same response) ✓ Including a qualitative open-ended question as an attention check ✓ Designing a short study (approximately 5 minutes) ✓ Avoiding using scales that have only "end" points labeled ✓ Repeating pertinent questions at the end of the study that make explicit the desired response and including a "Quit study" and "Contact researcher" option on each page	• MTurker inattention • Self-misrepresentation • Vulnerability to web robots (or "bots") • Perceived researcher unfairness

(continued)

Table 3 (continued)

Stage of Study	Recommendation	Implementation Guidelines	MTurk Challenge Addressed (From Table 2)
	6. Craft the MTurk task or Human Intelligence Task (HIT)	✓ Providing a detailed description that includes accurate estimated time commitment, what MTurkers will be asked to do, and compensation rules ✓ Avoiding cues that might provide MTurkers with signals about the study's aims or that might motivate MTurkers to engage in self-misrepresentation or exhibit greater social desirability bias (see online supplement Appendix H for a generic and customizable HIT post)	• Self-misrepresentation • MTurker social desirability bias
Implementation	7. Launch the study, monitor responses, and respond to concerns	✓ Conducting a pilot test with a minimum of 10 to 30 participants that includes an open-ended question requesting feedback ✓ Monitoring MTurker communities to gauge MTurkers' reactions to the study ✓ Responding promptly to any questions or concerns raised by participants	• Growth of MTurker communities • Perceived researcher unfairness
	8. Screen data	✓ Screening data in a timely manner using at least two or more tools (e.g., MTurker self-reports of response effort, answers to attention checks, response times, statistical tools that analyze answer-choice response patterns, IP addresses, and open-ended qualitative questions) to estimate likely percentage of unusable responses ✓ Adjusting number of participants to achieve desired sample size	• MTurker inattention • High attrition rates • Vulnerability to bots
	9. Approve or deny compensation for completed responses	✓ Approving or denying compensation for completed responses within 24 to 48 hours of the MTurker completing the study ✓ Specifying the reason for rejecting compensation	• Perceived researcher unfairness
Reporting	10. Report details to ensure transparency	✓ Reporting information regarding all procedures followed, decisions made, and results obtained during each stage of the study ✓ Providing all necessary data for future, secondary analyses (e.g., meta-analyses) of findings (i.e., demographics, means, standard deviations, and effect sizes) ✓ Reporting details regarding the posting of the HIT, qualifications used to restrict access to the HIT, and detailed sample characteristics ✓ Explaining all decisions regarding the use of attention checks and screening techniques, including the number of participants excluded for each, decisions regarding sampling from particular countries, measurement equivalence when testing non-native English speakers, and non-naivete ✓ Reporting detailed characteristics of the study, including information related to time commitment required and compensation provided	• MTurker inattention • High attrition rates • Inconsistent English language fluency • MTurker non-naivete • Perceived researcher unfairness

2. Decide qualifications used to screen MTurkers. Formulating study-appropriate protocols to screen MTurkers helps address threats posed by self-misrepresentation, inconsistent English language fluency, and MTurker non-naivete. First, to address self-misrepresentation, there is a need to be explicit about the qualifications (e.g., age, years of work experience, race) relevant for the study. Then, rather than explicitly listing desired qualifications, which can motivate self-misrepresentation, one can evaluate MTurkers using a screener study, pay everyone who participates, eliminate those who do not match desired criteria, and invite those who meet the qualifications/pass the screener to participate in the focal study (Chandler, Mueller, & Paolacci, 2014; Hydock, 2018; Siegel, Navarro, & Thomson, 2015; Wessling, Huber, & Netzer, 2017). This technique is especially useful when attempting to recruit unique populations (e.g., participants who identify as LGBTQ; Casey et al., 2017). Second, to address inconsistent English language fluency, one can determine a priori whether to consider only MTurkers from native-English-speaking countries (based on their internet protocol [IP] addresses), or to establish measurement equivalence across native and non-native English speakers (Feitosa, Joseph, & Newman, 2015). Finally, regarding MTurker non-naivete, there is a need to decide whether to use only highly qualified MTurkers (i.e., "Master Workers" who have considerable experience as an MTurker and therefore greater familiarity with common manipulations, attention check techniques, and experimental tasks and questions; Lovett, Bajaba, Lovett, & Simmering, 2018; Peer, Vosgerau, & Acquisti, 2014) or, alternatively, employ screening questions to gauge MTurker familiarity with research subject, stimuli, and if applicable, manipulations.

3. Establish required sample size. Many MTurker responses are unusable due to high attrition rates and MTurker inattention. Therefore, in addition to the sample size determined through a power analysis, it is useful to collect data from at least an additional 15% to 30% of MTurkers (Sprouse, 2011) to compensate for participant attrition and failure to pass attention and compliance checks (Barends & de Vries, 2019; Zhou & Fishbach, 2016).

4. Formulate compensation rules. Clear rules regarding compensation help address the threat posed by the challenge of perceived researcher unfairness. Higher MTurker pay is also linked to better performance on research tasks (Casey et al., 2017). The recommendation is to pay a fair wage in relation to the tasks required of the MTurker (Crump, McDonnell, & Gureckis, 2013), typically the minimum wage when drawing on samples from the United States (Buhrmester, Talaifar, & Gosling, 2018; Horton & Chilton, 2010; Litman, Robinson, & Rosenzweig, 2015; Liu & Sundar, 2018). In addition, researchers should decide a priori what criteria (if any) will be used to refuse payment to MTurkers (Fieseler, Bucher, & Hoffmann, 2017; Gleibs, 2017), and the schedule of payment. Moreover, codes of conduct, monitoring procedures, and penalties for fraudulent or untruthful reporting should be formulated as levying economic penalties for deceitful behavior can affect MTurkers' honesty (Brink, Eaton, Grenier, & Reffett, 2019). These norms should be made explicit and shared with participants in the consent form. As an additional resource, online supplement Appendix G includes a sample template of a consent form that can be customized for use in future MTurk research.

5. Design data collection tool used to gather responses. A well-designed data collection tool can help researchers address validity threats posed by the challenges of vulnerability to web robots, self-misrepresentation, MTurker inattention, and perceived researcher unfairness. We offer five recommendations. First, MTurkers should complete an informed consent form (Bederson & Quinn, 2011), which includes a "CAPTCHA" verification to thwart web robots—a "Completely Automated Public Turing Test to tell Computers and Humans Apart" that discerns human responses from web robots (von Ahn, Blum, Hopper, & Langford, 2003). This is done by having respondents correctly answer a set of challenges (e.g., identify pictures, type in words) to proceed. In addition, it is useful to include procedures designed to capture an MTurkers' IP address and use features that prevent the same MTurker from completing the study more than once (i.e., avoiding "ballot box stuffing"; Buhrmester et al., 2018; Chandler & Paolacci, 2017).

Second, it is useful to require MTurkers to provide their MTurk ID and maintain a reference database of past participants. This helps identify MTurkers who attempt self-misrepresentation to qualify for a particular study (Stewart et al., 2015).

Third, to address the threat posed by MTurker inattention, it is helpful to use attention checks. While more is preferable, a minimum of two such checks should be employed (Ramsey, Thompson, McKenzie, & Rosenbaum, 2016; Thomas & Clifford, 2017). Types of attention checks include instructed items that direct MTurkers to complete or abstain from a particular action, bogus items that ask MTurkers to answer obvious or ridiculous questions, self-reports of effort, and questions on which all or almost all respondents should provide the same response (Huang, Bowling, Liu, & Li, 2015). Specifically for MTurk, it is necessary to include at least one open-ended question as an attention check to help address both MTurker inattention and vulnerability to web robots (Dennis, Goodson, & Pearson, 2019). The use of such items does not negatively affect data quality as long as items used are specifically developed for this purpose, as opposed to being drawn from other sources or created ad hoc (Huang et al., 2015).

Fourth, designing short studies (i.e., no more than 5 minutes to complete) and avoiding using scales that have only the "end" points labeled (e.g., a Likert-type scale labeled only 1 = *strongly agree*, 5 = *strongly disagree*) can help minimize MTurker inattention (Goodman, Cryder, & Cheema, 2013; Hamby & Taylor, 2016).

Fifth, to gauge social desirability, especially in experimental designs, it is useful to repeat pertinent questions at the end of the study that make explicit the desired response, and contrast participant answers to the same questions as when presented earlier (De Quidt, Haushofer, & Roth, 2018). Finally, it is helpful to include a "Quit study" and "Contact researcher" option on each page of the study (as applicable) to allow MTurkers to exit the study or ask questions, thereby addressing the threat posed by the challenge of perceived researcher unfairness (Mason & Suri, 2012; Schulze, Seedorf, Geiger, Kaufmann, & Schader, 2011).

6. Craft the MTurk task or Human Intelligence Task (HIT). The last action of the planning stage is designing the HIT or job posting that will be seen by MTurkers. Because one of the main complaints by MTurkers is that the HIT description and instructions are unclear (Lovett et al., 2018; Schulze et al., 2011), the description should include details about the study, such as an accurate estimated time commitment, what MTurkers will be asked to do, and compensation rules (Zhou & Fishbach, 2016). At the same time, researchers have to be careful to avoid cues that might provide MTurkers with signals about the study's aims or that might motivate MTurkers to engage in self-misrepresentation or exhibit greater social desirability bias. As an additional resource, online supplement Appendix H includes a template for a HIT post that can be customized for use in future MTurk research.

Implementation Stage

7. Launch the study, monitor responses, and respond to concerns. To ensure study instructions are clear and to identify and rectify potential data-quality or programming problems before the data are collected, it is useful to conduct a pilot test with a minimum of 10 to 30 participants that includes an open-ended question requesting feedback (Kees, Berry, Burton, & Sheehan, 2017). Once the study is launched, researchers can monitor MTurker communities (e.g., Turker Nation, MTurk Crowd) to gauge MTurkers reactions to the study (if any), check if pertinent information is being shared, and respond promptly to any questions or concerns raised by participants (Barchard & Williams, 2008; Brawley & Pury, 2016; Deng, Joshi, & Galliers, 2016). Together, these steps help address the threat posed by the growth of MTurker communities and perceived researcher unfairness.

8. Screen data. Screening MTurk data in a timely manner helps estimate the likely percentage of unusable responses. This information can then be used to adjust the number of potential participants to achieve the required sample size. Unusable responses can usually be attributed either to careless or insufficient-effort responding (IER) or to fraudulent and duplicate efforts. General tools that can be used to screen data for careless responding or IER include MTurker self-reports of effort (e.g., self-reported carelessness, rushed responding, and skipping of instructions), answers to attention checks (e.g., directed questions), response times, and statistical tools that analyze answer choice response patterns (Wood, Harms, Lowman, & DeSimone, 2017).

MTurkers who score higher on self-reports of response effort or fail to comply with directed questions are more likely to have engaged in careless responding or IER (Berinsky, Margolis, & Sances, 2014; Maniaci & Rogge, 2014). Thus, their responses can be compared with those of other MTurkers before deciding to include or exclude them. When evaluating response times, a best practice is to exclude participants who complete the task in less than one or two seconds per item (Wood et al., 2017). Finally, the most effective statistical tools that can be employed include: (a) long-string index (in which participant response patterns in choosing the same response for multiple items are analyzed for frequency and length, and a threshold is developed based on the data to indicate potentially invalid responses; Hong, Steedle, & Cheng, 2020; Johnson, 2005; Maniaci & Rogge, 2014); and (b) within-session response consistency (which calculates the level of similarity in a participant's responses to items they have rated twice and excludes responses that score below 0.25; Wood et al., 2017). At least two of the aforementioned recommendations should be used to screen data (Buchanan & Scofield, 2018).

Regarding fraudulent or duplicate efforts, the most commonly used method is to examine IP addresses and delete duplicates. However, the growing popularity of virtual private servers that conceal the IP address of the device used to access the MTurk study are making it harder to rely solely on this screening procedure (Dennis et al., 2019). Furthermore, if multiple MTurkers use the same device (e.g., a laptop in a dorm room or a computer laboratory, a shared phone or tablet), their IP addresses will be the same, which can cause researchers to mistakenly omit legitimate responses. Accordingly, in addition to employing IP address screening (e.g., using software packages for R and Stata designed by Kennedy, Clifford, Burleigh, Jewell, & Waggoner, 2018), it is useful to examine the response to the open-ended attention check question included in the study (Dennis et al., 2019) before making the decision to include or omit a particular response. Overall, these steps help address threats posed by the challenges of MTurker inattention, vulnerability to web robots, high attrition rates, and English fluency.

9. Approve or deny compensation for completed responses. Based on data screening and using a priori rules, one can approve or deny compensation within 24 to 48 hours of the MTurker completing the study (Bederson & Quinn, 2011). Researchers can also specify the reason for rejecting compensation (Brawley & Pury, 2016; Gleibs, 2017). These steps help address the threat posed by the challenge of perceived researcher unfairness.

Reporting Stage

10. Report details to ensure transparency. There are growing calls in management and many other fields about the need for greater transparency regarding specific procedures, judgment calls, and decisions made during a study (Aguinis et al., 2018; Aguinis, Banks, Rogelberg, & Cascio, 2020; Aguinis & Solarino, 2019). These concerns are even more relevant for MTurk studies as participants are anonymous and often cannot be contacted for clarification. Accordingly, to address concerns about how different challenges may threaten validity of results obtained and conclusions reached when using MTurk (Hydock, 2018; Rouse, 2015), there is a need to clearly describe all steps (Thomas & Clifford, 2017; Zhou & Fishbach, 2016). First, studies should provide all necessary data for future, secondary analyses (e.g., meta-analyses) of their findings (i.e., demographic data, means, standard deviations, and effect sizes). In addition, there is a need to report details regarding the posting of the HIT (i.e., were data collected in one batch or multiple batches, was the HIT reposted), qualifications used to restrict access to the HIT (e.g., age, country of residence, Master Worker status), and detailed sample characteristics (e.g., gender, race/ethnicity, employment status, work experience, educational qualifications). Furthermore, it is necessary to report details regarding the use of attention checks and screening techniques, including the number of participants excluded for each (Cheung, Burns, Sinclair, & Sliter, 2017), as well as decisions regarding sampling from particular countries, measurement equivalence when testing non-native English speakers, and non-naivete (Chandler et al., 2014; Feitosa et al., 2015). To address ethical concerns, it is useful to provide detailed information related to time commitment required and compensation provided (Gleibs, 2017; Keith, Tay, & Harms, 2017).

In closing, our recommendations offer guidance for researchers using MTurk, journal editors and reviewers who evaluate submitted manuscripts, and consumers of research (i.e., other researchers, managers, consultants, policy makers) who wish to determine whether research using MTurk is sufficiently trustworthy. We hope our article will serve as a catalyst for more robust, reproducible, and trustworthy MTurk-based research in management and related fields.

ORCID iDs

Herman Aguinis https://orcid.org/0000-0002-3485-9484

Isabel Villamor https://orcid.org/0000-0003-0845-7016

Ravi S. Ramani https://orcid.org/0000-0003-3324-3765

References

Aguinis, H., Banks, G. C., Rogelberg, S. G., & Cascio, W. F. in press. Actionable recommendations for narrowing the science-practice gap in open science. *Organizational Behavior and Human Decision Processes*, 158: 27-35.

Aguinis, H., Ramani, R. S., & Alabduljader, N. 2018. What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, 12: 83-110.

Aguinis, H., Ramani, R. S., & Alabduljader, N. 2020. Best-practice recommendations for producers, evaluators, and users of methodological literature reviews. *Organizational Research Methods*. Advance online publication. doi:10.1177/1094428120943281

Aguinis, H., & Solarino, A. M. 2019. Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*, 40: 1291-1315.

Alonso, O., & Mizzaro, S. 2012. Using crowdsourcing for TREC relevance assessment. *Information Processing & Management*, 48: 1053-1066.

Antin, J., & Shaw, A. 2012. Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. In *ACM Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 2925-2934. New York: ACM.

Arechar, A. A., Gächter, S., & Molleman, L. 2018. Conducting interactive experiments online. *Experimental Economics*, 21: 99-131.

Bader, F., Baumeister, B., Berger, R., & Keuschnigg, M. in press. On the transportability of laboratory results. *Sociological Methods & Research*. Advance online publication. doi:10.1177/0049124119826151

Barchard, K. A., & Williams, J. 2008. Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods*, 40: 1111-1128.

Barends, A. J., & de Vries, R. E. 2019. Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences*, 143: 84-89.

Bederson, B. B., & Quinn, A. J. 2011. Web workers unite! Addressing challenges of online laborers. *ACM CHI Extended Abstracts on Human Factors in Computing Systems*, 2011: 97-106.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43: 800-813.

Bergvall-Kåreborn, B., & Howcroft, D. 2014. Amazon Mechanical Turk and the commodification of labor. *New Technology, Work, and Employment*, 29: 213-223.

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: 351-368.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. 2014. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58: 739-753.

Bernerth, J., & Aguinis, H. 2016. A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, 69: 229-283.

Brawley, A. M., & Pury, C. L. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior*, 54: 531-546.

Brink, W. D., Eaton, T. V., Grenier, J. H., & Reffett, A. 2019. Deterring unethical behavior in online labor markets. *Journal of Business Ethics*, 156: 71-88.

Buchanan, E. M., & Scofield, J. E. 2018. Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, 50: 1-11.

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. 2018. An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13: 149-154.

Bunge, E., Cook, H. M., Bond, M., Williamson, R. E., Cano, M., Barrera, A. Z., & . . . Muñoz, R. F. 2018. Comparing Amazon Mechanical Turk with unpaid internet resources in online clinical trials. *Internet Interventions*, 12: 68-73.

Callison-Burch, C., & Dredze, M. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT workshop on creating speech and language data with Amazon's Mechanical Turk*: 1-12. Stroudburg, PA: Association for Computational Linguistics.

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. 2017. Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, 7: 2158244017712774.

Casler, K., Bickel, L., & Hackett, E. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29: 2156-2160.

Chandler, J., Mueller, P., & Paolacci, G. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46: 112-130.

Chandler, J. J., & Paolacci, G. 2017. Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8: 500-508.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. 2015. Using non-naive participants can reduce effect sizes. *Psychological Science*, 26: 1131-1139.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. 2019. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51: 2022-2038.

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: 347-361.

Clifford, S., & Jerit, J. 2014. Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, 1: 120-131.

Colman, D. E., Vineyard, J., & Letzring, T. D. 2018. Exploring beyond simple demographic variables: Differences between traditional laboratory samples and crowdsourced online samples on the Big Five personality traits. *Personality and Individual Differences*, 133: 41-46.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, 8: e57410.

De Quidt, J., Haushofer, J., & Roth, C. 2018. Measuring and bounding experimenter demand. *American Economic Review*, 108: 3266-3302.

Deng, X., Joshi, K. D., & Galliers, R. D. 2016. The duality of empowerment and marginalization in microtask crowdsourcing: Giving voice to the less powerful through value sensitive design. *MIS Quarterly*, 40: 279-302.

Dennis, S. A., Goodson, B. M., & Pearson, C. 2019. *Virtual private servers and the limitations of IP-based screening procedures: Lessons from the MTurk quality crisis of 2018* (Publication No. 3233954). SSRN.

Feitosa, J., Joseph, D. L., & Newman, D. A. 2015. Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, 75: 47-52.

Fieseler, C., Bucher, E., & Hoffmann, C. P. 2017. Unfairness by design? The perceived fairness of digital labor on crowdworking platforms. *Journal of Business Ethics*, 156: 987-1005.

Gleibs, I. H. 2017. Are all "research fields" equal? Rethinking practice for the use of data from crowdsourcing market places. *Behavior Research Methods*, 49: 1333-1342.

Goodman, J. K., Cryder, C. E., & Cheema, A. 2013. Data collection in a flat world: The advantages and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26: 213-224.

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: 912-932.

Heer, J., & Bostock, M. 2010. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *ACM Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 203-212. New York: ACM.

Hong, M., Steedle, J. T., & Cheng, Y. 2020. Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, 80: 312-345.

Horton, J. J., & Chilton, L. B. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the ACM Conference on Electronic Commerce*: 209-218. New York: ACM.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. 2015. Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, 30: 299-311.

Hydock, C. 2018. Assessing and overcoming participant dishonesty in online data collection. *Behavior Research Methods*, 50: 1563-1567.

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

Kan, I. P., & Drummey, A. B. 2018. Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce. *Computers in Human Behavior*, 83: 243-253.
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: 141-155.
Keith, M. G., Tay, L., & Harms, P. D. 2017. Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in Psychology*, 8: 1-19.
Kennedy, R., Clifford, S., Burleigh, T., Jewell, R., & Waggoner, P. 2018. *The shape of and solutions to the MTurk quality crisis* (Publication No. 3272468). SSRN.
Landers, R. N., & Behrend, T. S. 2015. An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology*, 8: 142-164.
Levay, K. E., Freese, J., & Druckman, J. N. 2016. The demographic and political composition of Mechanical Turk samples. *Sage Open*, 6: 2158244016636433.
Litman, L., Robinson, J., & Rosenzweig, C. 2015. The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, 47: 519-528.
Liu, B., & Sundar, S. S. 2018. Microworkers as research participants: Does underpaying Turkers lead to cognitive dissonance? *Computers in Human Behavior*, 24: 89-101.
Lovett, M., Bajaba, S., Lovett, M., & Simmering, M. J. 2018. Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of Amazon’s Mechanical Turk Masters. *Applied Psychology*, 67: 339-366.
Maniaci, M. R., & Rogge, R. D. 2014. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48: 61-83.
Mason, W., & Suri, S. 2012. Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods*, 44: 1-23.
Mummolo, J., & Peterson, E. 2019. Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, 113: 517-529.
Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. 2016. Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PLOS ONE*, 11: e0157732.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5: 411-419.
Paré, G., Trudel, M. C., Jaana, M., & Kitsiou, S. 2015. Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52: 183-199.
Pearl, J., & Bareinboim, E. 2014. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29: 579-595.
Peer, E., Vosgerau, J., & Acquisti, A. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46: 1023-1031.
Porter, C. O., Outlaw, R., Gale, J. P., & Cho, T. S. 2019. The use of online panel data in management research: A review and recommendations. *Journal of Management*, 45: 319-344.
Ramsey, S. R., Thompson, K. L., McKenzie, M., & Rosenbaum, A. 2016. Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior*, 58: 354-360.
Rouse, S. V. 2015. A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, 43: 304-307.
Schulze, T., Seedorf, S., Geiger, D., Kaufmann, N., & Schader, M. 2011. Exploring task properties in crowdsourcing: An empirical study on Mechanical Turk. In *ECIS Proceedings*: 122. http://aisel.aisnet.org/ecis2011/122.
Siegel, J. T., Navarro, M. A., & Thomson, A. L. 2015. The impact of overtly listing eligibility requirements on MTurk: An investigation involving organ donation, recruitment scripts, and feelings of elevation. *Social Science & Medicine*, 142: 256-260.
Sprouse, J. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 43: 155-167.
Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. 2015. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, 10: 479-491.
Stritch, J. M., Pedersen, M. J., & Taggart, G. 2017. The opportunities and limitations of using Mechanical Turk (MTurk) in public administration and management scholarship. *International Public Management Journal*, 20: 489-511.

Summerville, A., & Chartier, C. R. 2013. Pseudo-dyadic “interaction” on Amazon’s Mechanical Turk. *Behavior Research Methods*, 45: 116-124.

Thomas, K. A., & Clifford, S. 2017. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, 77: 184-197.

Tosti-Kharas, J., & Conley, C. 2016. Coding psychological constructs in text using Mechanical Turk: A reliable, accurate, and efficient alternative. *Frontiers in Psychology*, 7: 741.

von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. 2003. CAPTCHA: Using hard AI problems for security. *Lecture Notes in Computer Science*, 2656: 294-311.

Walter, S. L., Seibert, S. E., Goering, D., & O’Boyle, E. H. 2019. A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology*, 34: 425-452.

Weinberg, J., Freese, J., & McElhattan, D. 2014. Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample. *Sociological Science*, 1: 292-310.

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

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. 2017. Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, 8: 454-464.

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

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