Journal of the Midwest Association for Information Systems

Volume 2019 | Issue 2 Article 2

Date: 07-31-2019

Don't Get Lost in the Crowd: Best Practices for Using Amazon's Mechanical Turk in Behavioral Research

Jacob A. Young
Bradley University, jayoung@fsmail.bradley.edu

Kristie M. Young
Illinois State University, kmyou3@ilstu.edu

Abstract

The use of Amazon's Mechanical Turk (MTurk) to conduct academic research has steadily grown since its inception in 2005. The ability to control every aspect of a study, from sampling to collection, is extremely appealing to researchers. Unfortunately, the additional control offered through MTurk can also lead to poor data quality if researchers are not careful. Despite research on various aspects of data quality, participant compensation, and participant demographics, the academic literature still lacks a practical guide to the effective use of settings and features in MTurk for survey and experimental research. Therefore, the purpose of this tutorial is to provide researchers with a recommended set of best practices to follow before, during, and after collecting data via MTurk to ensure that responses are of the highest possible quality. We also recommend that editors and reviewers place more emphasis on the collection methods employed by researchers, rather than assume that all samples collected using a given online platform are of equal quality. We also recommend that editors and reviewers place more emphasis on the collection methods employed by researchers, rather than assuming that all samples collected using a given online platform are of equal quality.

Keywords: Mechanical Turk, sampling, survey research, experimental research

DOI:10.17705/3jmwa.000050

1. Introduction

Amazon Mechanical Turk (MTurk), an online crowdsourcing platform, has emerged as an attractive data collection method for both survey and experimental research (Buhrmester, Talaifar, & Gosling, 2018). Even though the use of MTurk has increased, authors still find themselves forced to defend the quality of data collected on MTurk to reviewers and editors. Lowry, D'Arcy, Hammer, and Moody (2016) had this to say on the issue:

> *A pattern has taken hold in which traditional organizational researchers, reviewers, and editors are quick to misconstrue and reject new methods while defending the "best practices" of paper surveys, which have been the methodology of choice for several decades. Although organizations themselves have implemented significant innovations, the published research on organizations has not undertaken innovation to the same degree. Traditionalists and the researchers who make up the reviewing system in the organization science and information systems (IS) fields are quick to downplay the legitimacy of new theories and methods, but they fail to apply the same level of scrutiny to their own traditions. This thwarts scientific progress. (Lowry et al., 2016, p. 233).*

Critiquing the quality of all data during the review process is certainly important. However, we argue that the efficacy of MTurk as a research tool, as opposed to more widely accepted online panel services (e.g. Qualtrics, SurveyMonkey, Turkprime) or traditional paper surveys, should be judged based upon the qualification methodology employed by the researcher rather than the collection media itself (Landers & Behrend, 2015; Roulin, 2015). We argue that the use of all online panels, where researchers pay for a study instrument to be administered to a group of prequalified participants, reduces the validity and generalizability of behavioral research. The inability to confirm or even fully describe the procedures used to develop and validate a given sample is a major disadvantage to the use of these services because it forces authors, reviewers, and editors to blindly accept the quality of the panel. Therefore, we argue that sample reliability and study generalizability is greatly improved if researchers are required to document how they qualified their subjects instead of accepting panels qualified by such services.

Despite promising research on various aspects of MTurk, such as data quality (Behrend, Sharek, Meade, & Wiebe, 2011; Buhrmester, Kwang, & Gosling, 2011; Landers & Behrend, 2015; Paolacci, Chandler, & Ipeirotis, 2010; Peer, Vosgerau, & Acquisti, 2014; Shapiro, Chandler, & Mueller, 2013; Sprouse, 2011; Steelman, Hammer, & Limayem, 2014), participant compensation (Chandler, Paolacci, & Mueller, 2013; Deng & Joshi, 2016; Goodman, Cryder, & Cheema, 2013; Horton & Chilton, 2010; Kraut et al., 2004; Mason & Suri, 2012; Mason & Watts, 2009) participant diversity (Behrend et al., 2011; Buhrmester et al., 2011; Kaufmann & Veit, 2011; Kraut et al., 2004; Mason & Suri, 2012; Paolacci et al., 2010; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010), and successful replications (Berinsky, Huber, & Lenz, 2012; Crump, McDonnell, & Gureckis, 2013; Horton, Rand, & Zeckhauser, 2011) the academic literature lacks a practical guide to the effective use of MTurk for survey and experimental research. Thus, the purpose of this tutorial paper is to provide behavioral researchers with a suggested set of best practices to follow when employing MTurk to ensure that future research is based on high-quality data. Due to our specific focus on MTurk, we only mention traditional best practices (e.g., Kerlinger & Lee, 2000; Pedhazur & Schmelkin, 1991; Shadish, Cook, & Campbell, 2002) when describing how to apply them on MTurk. Therefore, researchers must ensure that proper statistical and experimental procedures have been followed when using MTurk, just as they should with any other sampling method.

Although there are many possible uses for MTurk, we limited our paper to its use in survey and experimental research. Our suggested best practices build upon discussions found in prior literature (Cheung, Burns, Sinclair, & Sliter, 2017; Jia, Reich, & Jia, 2016; Jia, Steelman, Reich, & Jia, 2017; Lowry et al., 2016) in addition to knowledge gained through our personal use of MTurk. Cheung et al. (2017) discuss methodological concerns with MTurk and provide general recommendations based on the work of Shadish, Cook, & Campbell (2002). However, they provide little guidance on exactly how to address these concerns when using MTurk. Similarly, Jia et al. (2017) provide a table of recommendations with brief rationales, but also lack specific instruction on how to follow these recommendations when using MTurk. Further, Jia et al. (2017, p. 309) contend that MTurk is only suitable for research that can be "generalized to a variety of users and technologies" and samples populations with "diverse individual cognition." We contend that if authors properly follow the best practices we outline in this paper, all types of behavioral research can be conducted on MTurk without diminishing data quality, especially when compared to other online sampling methods.

Another issue is that due to page limits, authors tend to provide brief explanations of their data collection procedures (Lowry et al., 2016). For example, in a study of industrial-organizational psychology journals, approximately one-third of the articles did not include any information on quality control measures (Cheung et al., 2017). Because of this, determining whether the sampling methodology was adequately implemented and examined during the review process is often difficult, regardless of the sampling method employed. Therefore, assessing whether our suggested best practices are being followed when collecting data via MTurk is difficult, and evaluating the techniques employed in studies that have been rejected and remain unpublished is impossible.

To address these issues, we discuss specific options and settings available in MTurk to employ best practices. First, we provide an overview of MTurk. Second, we propose best practices for working with Workers on MTurk. Third, we outline the suggested best practices in this tutorial with respect to phases of the sampling process: before, during, and after data collection. We conclude the tutorial by providing recommendations for authors, editors, and reviewers to aid in the assessment and reporting of data quality and collection procedures when using MTurk. We also compare and contrast our recommendations with those of Cheung et al. (2017), Jia et al. (2017), and Lowry et al. (2016) throughout our paper. We have provided appendices to help researchers outline the expectations and instructions to participants of studies conducted using MTurk.

2. Overview of Amazon Mechanical Turk

Crowdsourcing has been defined as "the paid recruitment of an online, independent global workforce for the objective of working on a specifically defined task or set of tasks" (Behrend et al., 2011, p. 801). Amazon Mechanical Turk is a crowdsourcing platform that serves as an online marketplace for individuals and businesses, referred to as *Requesters*, to hire independent contractors, referred to as *Workers*, to remotely perform a wide variety of jobs, referred to as *Human Intelligence Tasks* (HITs). Requesters choose the payment amount and participant qualifications. Requesters review work and determine if it should be accepted or rejected, or if a bonus payment is appropriate. Workers' reputations are indicated by their HIT acceptance rate, while Requesters' reputations are based on opinions shared by Workers on external websites. Behavioral researchers are most likely to use MTurk to solicit participants for surveys and experiments, and then conduct the study on other online research platforms, such as Qualtrics or SurveyMonkey. For an excellent introduction to MTurk and its uses in behavioral research, see Mason & Suri (2012).

2.1. Benefits of MTurk

Although Amazon does not reveal user information, several studies have reported on the characteristics of Workers and Requesters. Ipeirotis (2010b) determined that when compared to Internet users in general, Workers tend to be younger, mainly female, and have less income. It is estimated that there are currently over 100,000 users on MTurk, with at least 2,000 actives at any given time (Difallah, Filatova, & Ipeirotis, 2018). Demographic data for certain date ranges can also be obtained from Mechanical Turk Tracker (http://mturk-tracker.com) (Difallah, Catasta, Demartini, Ipeirotis, & Cudré-Mauroux, 2015; Ipeirotis, 2010a). As shown in Table 1, several studies have identified numerous benefits of using MTurk over other primary data sources. Further, MTurk is particularly useful in behavioral research (Behrend et al., 2011; Goodman et al., 2013), allowing surveys and experiments to be conducted online without sacrificing quality (Briones & Benham, 2017; Mason & Watts, 2009; Rogstadius et al., 2011; Sprouse, 2011). While Goodman et al. (2013) hypothesized that MTurk participants might disregard instructions if it is likely to lead to a higher payment, the study found that cheating was significantly reduced from 40.1 to 27.2 percent simply by asking MTurk participants to answer honestly.

Ultimately, MTurk provides researchers with greater control and flexibility at less expense than other online panel providers. MTurk's pricing is far more transparent in that Amazon's base fee is a percentage of the amount paid directly to Worker(s) for completing a HIT (20 percent for batches with fewer than ten assignments and 40 percent for batches with ten or more assignments) (Amazon Mechanical Turk, n.d.-b), where an assignment is referring to one completion of the HIT. Other online panel providers typically charge researchers a flat fee per respondent. Unfortunately, pricing using the flat fee approach is indicative of the challenge in obtaining a sample of the desired population rather than the actual payment made to each respondent, which obfuscates the sampling methodology. For example, researchers might be quoted a cost of $50 per respondent to sample a niche target population with a short 10-minute survey, yet only $5 of that fee is paid to each respondent. The remaining $45 cost is incurred by the online panel provider in the recruitment and identification of the sample. Understandably, the online panel services do not want to reveal their internal cost structure, but the inability of researchers to report the true amount paid to respondents or the sample recruitment procedure used by the online panel service is problematic.

Category	References
Cost	Chandler et al., 2013; Goodman et al., 2013; Horton & Chilton, 2010; Kraut et al., 2004; Mason & Suri, 2012; Mason & Watts, 2009
Subject Pool Access	Behrend et al., 2011; Chandler et al., 2013; Goodman et al., 2013; Kraut et al., 2004; Lowry et al., 2016; Mason & Suri, 2012; Mason & Watts, 2009; Shapiro et al., 2013; Stewart et al., 2015
Subject Pool Diversity	Behrend et al., 2011; Buhrmester et al., 2011; Difallah et al., 2018; Kaufmann & Veit, 2011; Kraut et al., 2004; Lowry et al., 2016; Mason & Suri, 2012; Paolacci et al., 2010; Ross et al., 2010
Speed	Chandler et al., 2013; Goodman et al., 2013; Horton & Chilton, 2010; Lowry et al., 2016; Mason & Watts, 2009
Flexibility	Chandler et al., 2013; Kraut et al., 2004; Lowry et al., 2016; Mason & Watts, 2009
Attentiveness	Hauser & Schwarz, 2016
Anonymity	Chandler et al., 2013; Shapiro et al., 2013

Table 1. Advantages of using MTurk

2.2. Common Criticisms of MTurk

Some of the common criticisms of MTurk revolve around data verification, self-selection bias, and its appropriateness for sampling certain target populations. While all researchers should strive for perfect generalizability and validity, every study has its limitations. We contend that the control that researchers have when qualifying participants on MTurk is a substantial advantage over other online sampling methods.

Some of the benefits of online research might also negatively affect studies conducted using MTurk. For example, anonymity can certainly be beneficial to participants and reduce social desirability bias, but complete anonymity prevents researchers from verifying self-reported data (Cheung et al., 2017; Jia et al., 2017). Encouragingly, Rand (2012) found that most subjects answered reliably to demographic questions on MTurk. Unfortunately, one of the common misconceptions of MTurk is Workers falsely claiming to be residents of the United States (Jia et al., 2017). Previously, citizens of the United States were only required to provide either a social security number or an individual tax identification number upon reaching a certain level of earnings and international Workers were unable to perform any HITs without providing the necessary information found on IRS Form W-8BEN. Now, Amazon requires all Workers to provide valid taxpayer identification information when registering with Amazon Payments before they are permitted to complete a single HIT. This is explained in the frequently asked questions related to tax information on MTurk. Under “Tax Information for US Residents”, the answer to “Why am I asked to register with Amazon Payments?” states:

> *An Amazon Payments account allows you to transfer Amazon Mechanical Turk earnings to your bank account. We also require U.S. Workers to provide valid taxpayer identification information when registering with Amazon Payments. You must create an Amazon Payments account to work on HITs and your earnings may be subject to tax reporting with the Internal Revenue Service (IRS). To learn more, click here. (Amazon Mechanical Turk, n.d.-c)*

Under “Tax Information for Non-US Residents”, the answer to “Why am I asked to provide my tax information?” states:

> *We require Workers to provide valid taxpayer identification information in order to comply with U.S. tax reporting regulations governed by the U.S. tax authority (Internal Revenue Service or "IRS"). The tax information interview collects the information needed to complete an IRS tax form (e.g. IRS Form W-8) which will be used to certify your non-U.S. status, determine if your earnings are subject to IRS reporting, and the rate of U.S. tax withholding (if any) applicable to your earnings. (Amazon Mechanical Turk, n.d.-c)*

Amazon’s increased scrutiny for all new Worker accounts to address early issues with work performed by international participants might lead to a less diverse subject pool for studies requiring an international sample, but they have addressed much of the early criticism of MTurk with respect to sampling populations located in the United States.

Concerns have also been raised regarding self-selection bias (Cheung et al., 2017; Jia et al., 2017). We agree in the sense that there is no way to compel people to participate since everyone has autonomy. Yet, reduced verifiability and self-selection bias are potential issues true of all online samples and not limited to MTurk. Self-selection by ineligible participants can be mitigated by following our suggested best practices. Since these concerns exist for all online panels, we argue that if other platforms that provide far less control are considered acceptable, a study properly conducted on MTurk should be, as well. Jia et al. (2017) also discuss when MTurk samples or organizational samples are appropriate. We feel that MTurk has wider applications than what they suggest. We agree that when the topic of interest is narrow and specific to an organization, then an organizational sample is necessary. However, MTurk's extensive pool of potential participants and researcher control make it more advantageous than traditional sampling methods for studies intended to be generalizable to the population at large.

2.3. Dangers of Naïve Use of MTurk

The attractive benefits of MTurk introduce additional burdens that researchers must properly address to collect quality data. Due to its short existence and the relative ease of publishing HITs, researchers with little to no experience with MTurk might be unaware of potential data quality issues and how to mitigate them prior to data collection. In the following sections, we outline our suggested practices for maintaining a healthy relationship with Workers, as well as methods researchers should employ before, during, and after data collection on MTurk to ensure collection of the highest quality data possible. The ordering of the suggested practices is intended to follow the stages in the research process as best possible, though researchers should be aware that some of the practices can and should apply to multiple aspects of study design, data collection, and analysis. Therefore, we highly encourage readers to fully read and understand all the best practices before collecting data on MTurk.

3. Best Practices for Working with Workers

In addition to ensuring data quality, it is critical for researchers to maintain a symbiotic relationship with participants. Treating workers with respect and dignity preserves MTurk and other platforms as acceptable sources of participants for conducting research. Gleibs (2017) reminded researchers of the importance of maintaining ethical treatment while using crowdsourcing services. The best practices suggested in this section, summarized in Table 2, can help researchers maintain this vital relationship when utilizing MTurk.

Number	Best Practice	How to Implement
3.1	Protect Study Integrity and Reputation	• Be aware of your reputation • Resolve any issues quickly
3.2	Provide Clear Expectations and Instructions	• State expectations regarding attentiveness, time commitment, and compensation • Include any study-specific restrictions that you expect Workers to follow
3.3	Provide Contact Information	• Provide Workers with an email address that will be monitored during data collection • Allow Workers to provide feedback after completing the study
3.4	Be Fair and Consistent	• Set payment at or above minimum wage • Establish an objective rubric for submissions • Include a statement in the HIT description of how work will be assessed
3.5	Maintain Worker Confidentiality and Anonymity	• Protect participant information • Only collect anonymous MTurk Worker IDs

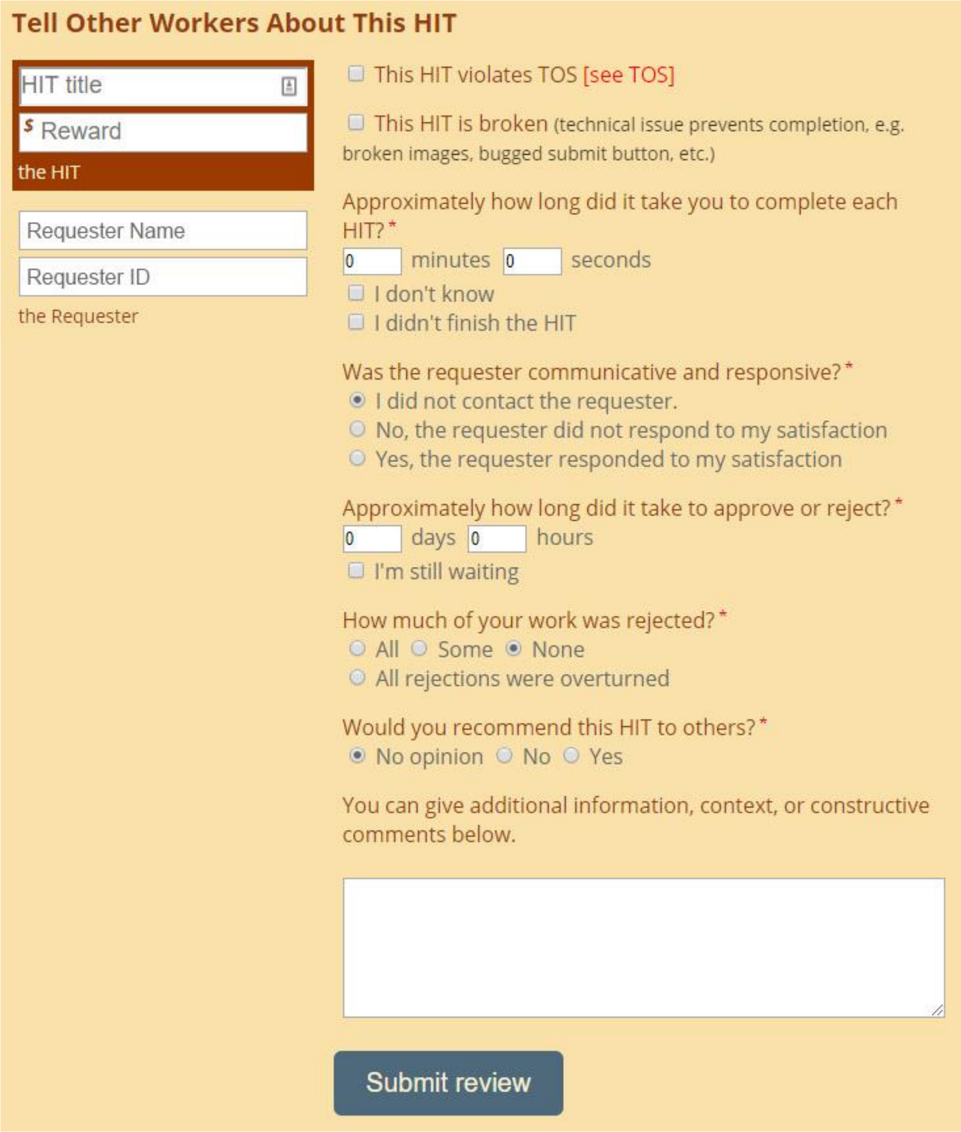
Table 2. Best Practices for Working with Workers

3.1. Protect Study Integrity & Reputation

We believe that effective use of MTurk requires obtaining accounts on multiple websites to protect the integrity of studies, as well as one's own reputation among MTurk Workers (Mason & Suri, 2012; Paolacci et al., 2010). In this section, we discuss their benefits and the best practices to follow when using these additional accounts.

Workers, who commonly refer to themselves as "Turkers," often post reviews of Requesters on MTurk review websites, such as Turkopticon (http://turkopticon.ucsd.edu/) and Turker Nation (http://www.turkernation.com) (Cheung et al., 2017; Jia et al., 2017; Mason & Suri, 2012). The use of these outlets might result in study-specific information, such as the location and answers to attention and manipulation checks, being shared with potential study participants, possibly invalidating the results of the study. Chandler, Mueller, & Paolacci (2014) concluded that cross-talk about study content among Workers was not a major problem. However, prohibiting participants from discussing the study on public forums and monitoring these websites during collection is still important to ensure that such disclosures have not compromised the integrity of the study.

The first version of Turkopticon (https://turkopticon.ucsd.edu) allows Workers to rate a Requester's "communicativity," "generosity," "fairness" and "promptness" on a scale of one to five, as well as submit detailed comments about their participation in a given HIT. The beta version of Turkopticon 2 (https://turkopticon.info/) has been modified to focus ratings on individual HITs rather than aggregate all ratings for each Requester. The rating criteria has also evolved to include items related to terms of service violations, technical issues, completion time, approval/rejection time, and whether the Worker would recommend the HIT to others. The HIT review form of Turkopticon 2 can be seen in Figure 1.



Tell Other Workers About This HIT

HIT title

$ Reward

the HIT

Requester Name

Requester ID

the Requester

□ This HIT violates TOS [see TOS]

□ This HIT is broken (technical issue prevents completion, e.g. broken images, bugged submit button, etc.)

Approximately how long did it take you to complete each HIT?*

0 minutes 0 seconds

□ I don't know

□ I didn't finish the HIT

Was the requester communicative and responsive?*

● I did not contact the requester.

○ No, the requester did not respond to my satisfaction

○ Yes, the requester responded to my satisfaction

Approximately how long did it take to approve or reject?*

0 days 0 hours

□ I'm still waiting

How much of your work was rejected?*

○ All ○ Some ● None

○ All rejections were overturned

Would you recommend this HIT to others?*

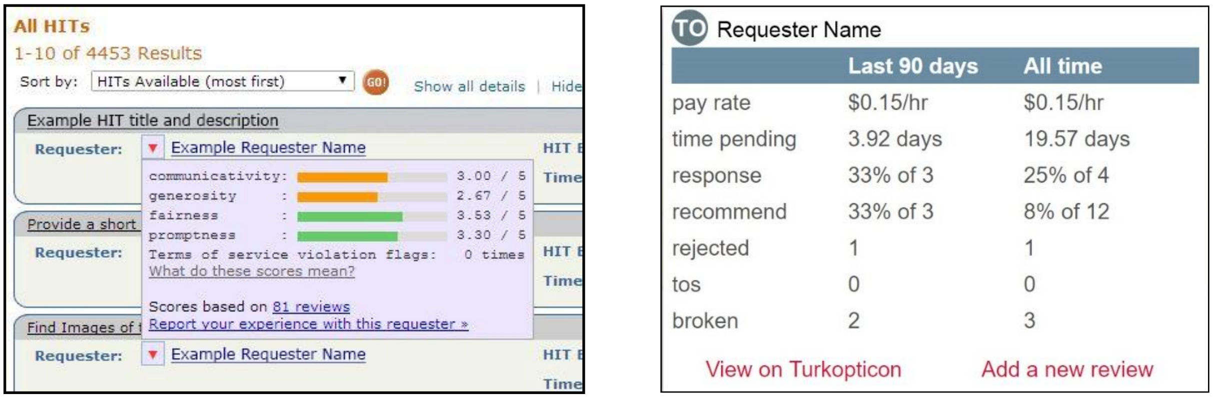
● No opinion ○ No ○ Yes

You can give additional information, context, or constructive comments below.

Submit review

Figure 1. Turkopticon 2 HIT Review Form

In addition to the Turkopticon websites, Requester reputation ratings are readily available to potential Workers who are using Internet browser plugins or have manually installed scripts (https://turkopticon.info/install). If the plugin or script is installed, Workers can quickly gain insight on the Requester's reputation in the Turkopticon community while browsing available HITs on MTurk. This information is provided by a pop-up box that can be accessed simply by hovering the mouse cursor over the small icon inserted in front of the Requester’s name for each HIT. A side-by-side comparison of Requester ratings as seen on MTurk using browser scripts for the original Turkopticon and beta version of Turkopticon 2 is provided in Figure 2.



```
communicativity:  3.00 / 5
generosity     :  2.67 / 5
fairness       :  3.53 / 5
promptness     :  3.30 / 5
Terms of service violation flags:   0 times
What do these scores mean?

Scores based on 81 reviews
Report your experience with this requester »
```

TO Requester Name

	Last 90 days	All time
pay rate	$0.15/hr	$0.15/hr
time pending	3.92 days	19.57 days
response	33% of 3	25% of 4
recommend	33% of 3	8% of 12
rejected	1	1
tos	0	0
broken	2	3

Figure 2. Comparison of Old and New Turkopticon Ratings as Viewed on MTurk

Requesters (based upon the original Turkopitcon) and HITs (based upon Turkopticon 2) with a poor reputation among Workers might struggle to attract study participants, and those who do elect to participate might not provide reliable data. Therefore, Requesters should be aware of their reputation and strive to resolve any issues Workers might have as reasonably and swiftly as possible. Reviewing valuable feedback from Workers can also help researchers improve future HITs and their standing in the MTurk community. For those who discover that they have poor reputations on these services, we recommend following our best practices under a new Requester account. This will provide a clean slate and allow the researcher to build a positive reputation over time. Aside from that, we do not encourage researchers to create new accounts unless compelling justifications can be given. The goal of this paper is for researchers to adopt best practices so MTurk will remain a mutually beneficial research platform. Repeatedly creating new accounts to avoid maintaining a poor Requester reputation is unethical and counter to the spirit of our recommendations.

3.2. Provide Clear Expectations and Instructions

Researchers should ensure that they have provided detailed expectations in the HIT description for potential participants to review on MTurk. They should also be upfront about compensation and time required (Paolacci et al., 2010) and notify participants that they will be removed for inattentiveness (Jia et al., 2017). Some have suggested that stating the scientific importance of a study might reduce participant inattentiveness (Fleischer, Mead, & Huang, 2015; Goodman et al., 2013). However, researchers should make sure that study instructions do not invalidate responses by priming participants (Cheung et al., 2017). Further, restating these expectations again on the research platform being used prior to the participants’ commencement of the study is always wise.

Despite the fact that MTurk Workers conduct their work in an unsupervised and uncontrolled environment, research has shown that Workers will respond to specific instructions that restrict certain behavior, such as looking up answers on the Internet (Goodman et al., 2013). Cheung et al. (2017) advise asking participants to reduce extraneous factors by using a certain Web browser or finding a quiet place to complete the task. Providing clear directions for acceptance is also important because there are many tasks on MTurk where priming is not a concern. For example, approximately 13% of submitted HITs are returned, giving Workers an opportunity to improve their work and have it accepted (Hara et al., 2018). However, resubmitting work is not an option for surveys and experiments since it would invalidate the results. Therefore, we suggest that researchers clearly outline the expectations and instructions for participants to improve the likelihood of achieving acceptable results. We have provided recommended language in Appendices A and B. We have also provided a supplementary file that includes alternate code to use for the HIT expectations on MTurk which prevents access to the study link until the HIT has been accepted, as shown in Appendix C.

3.3. Provide Contact Information

Providing direct contact information to potential participants prior to the commencement of a study is always a good practice and likely required by institutional review boards (IRBs). This should be done within the HIT instructions. Providing Workers with an email address that is associated with an institution or research organization is likely to increase the study’s legitimacy. As we discuss in more detail below, researchers should also be available during the data collection process, because some Workers will email the researchers directly with questions, concerns, or to report technical issues. Also, email clients might filter messages sent to Requesters via the MTurk messaging system into spam folders. Therefore, researchers should be sure to monitor the email address associated with their MTurk Requester account during the data collection process and resolve any issues as quickly as possible. We also recommend that researchers include open-response questions for Workers to provide HIT-related feedback within the study instrument (Mason & Suri, 2012). Such feedback often pertains to confusing directions and issues experienced with the functionality of the instrument.

3.4. Be Fair and Consistent

Offering compensation relative to the task length for a given HIT has been shown to impact participation from MTurk Workers and reasonable compensation can be expected to yield quality data (Buhrmester et al., 2011). Although some MTurk Workers will accept HITs for little compensation, and Requesters are not bound by minimum wage laws since Workers are considered independent contractors, this is considered poor practice on ethical grounds. Additionally, low compensation is likely to increase data collection time and can negatively impact Requester reputation (Mason & Suri, 2012). Thus, researchers should ensure that they fairly compensate Workers for the time spent participating in the study (Jia et al., 2017; Lowry et al., 2016). At a minimum, we suggest that compensation be set at or above the hourly minimum wage in relation to the anticipated length of time to complete the HIT. Since some participants will take longer than others, we recommend that the minimum rate be based upon the completion time for the 75th percentile from a pilot study. For example, based upon the current minimum wage in the United States of $7.25, a survey expected to take most participants approximately 20 minutes to complete (i.e., based upon completion times obtained during pilot testing) should pay participants approximately $2.50. This is not only ethical but has also been shown to be a factor for participant motivation (Deng & Joshi, 2016; Kaufmann & Veit, 2011) and can result in improved data quality (Buhrmester et al., 2011).

The most effective metric for determining Worker quality is the HIT acceptance rate. Fairness and consistency when approving and rejecting submitted work for HITs are critical. Adhering to community norms when rejecting work and explaining why the work was rejected is also important (Paolacci et al., 2010). Approving all submissions without assessing work quality increases data collection costs and reduces the effectiveness of the metric for other Requesters, whereas rejecting every instance of questionable work is likely to reduce the Researcher's reputation among Workers. Therefore, researchers must take reasonable steps to maintain a delicate balance between approval and rejection that is appropriate for the interests of both parties.

Since the HIT acceptance rate is critical to assessing Worker quality and the rejection of work often results in negative Requester reviews, we recommend that researchers be proactive by establishing clear criteria for reviewing work prior to publishing a HIT on MTurk. Our suggested approach is for researchers to establish an objective rubric for poor, marginal, acceptable, and excellent submissions based upon the requirements and expectations for the study in question. Researchers can then assess the standards for a given study by collecting pilot batches.

Once the quality of a submission has been determined, we recommend that researchers refer to the matrix provided in Table 3 to determine whether to accept the work, whether to provide a bonus to the Worker, as well as whether additional communication with the Worker is warranted. Doing so will allow for a more objective and efficient work approval process.

Work Quality	Accept Work?	Provide Bonus?	Send Message?
Poor	No	No	Yes
Marginal	No	Yes	Yes
Acceptable	Yes	No	No
Excellent	Yes	Yes	Yes

Table 3. Work Approval Matrix

Poor data would consist of submissions that clearly indicate the Worker did not put forth an honest attempt. For example, providing the same response for every question in a 100-item survey indicates insufficient effort. In such cases, researchers should reject the work and send Workers an explanation for why their work was rejected. We believe that this is the best way for researchers to preserve their Requester reputation, while also maintaining the integrity of the HIT acceptance rate. However, if the work involves completing a survey or participating in an experiment, we recommend that these explanations be given in general terms to protect the integrity of the study.

Marginal data consists of submissions that appear to be honest attempts yet fail to meet the stated expectations for acceptable work. For example, submissions that fail an unacceptable number of attention check questions would be considered marginal. Researchers should pay special attention to how they handle poor to marginal submissions. Our suggested approach for marginal submissions is to reject the work, provide the Worker with a detailed explanation outlining the reason(s) for doing so, and provide a bonus payment to compensate the Worker for taking the time to participate in the study. Ideally, we recommend that researchers provide a bonus amount equivalent to accepted work to Workers who spent the expected amount of time participating in the study. The use of a bonus payment simply serves as compensation for time spent producing work of marginal quality, while also preserving the integrity of the HIT acceptance rate as a measure of Worker quality. This might appear to reward Workers for rejected work, but qualitative responses from Workers have indicated that they would prefer to preserve a high acceptance rate rather than receive a monetary bonus since poor acceptance rates limit the HIT opportunities available to them in the future.

Obviously, researchers should accept data that meet the criteria for acceptable and excellent submissions. However, if possible, we suggest that work that meets the criteria for excellent submissions also be rewarded with additional compensation through bonus payments. In addition to meeting the standard for acceptable work, an excellent submission might also include extensive qualitative information related to the study's context or feedback on the behavior of the study instrument. Sending a message that thanks them for their excellent submission and the use of a bonus payment provides positive reinforcement to the Worker and shows that the researchers appreciate the Worker's thoughtful participation in the study. These small gestures help preserve the number of quality respondents available to participate in future research conducted on MTurk.

Researchers would be wise to include a statement in the HIT description of how they will assess work. For example, the following statements would explain the suggested method: "We will review work within **[X]** hours. Honest, attentive, and complete responses will be accepted. Your work will be rejected if it does not satisfy our quality standards. If your work is rejected, you will be compensated for your time through a bonus payment." This informs potential participants that Workers who submit honest attempts will always be compensated for their time. Adopting these suggested practices will help researchers protect their reputation among Workers while also maintaining the HIT acceptance rate as a reliable measure of Worker quality.

3.5. Maintain Worker Confidentiality and Anonymity

Before granting approval for a proposed study, IRBs often require assurances from researchers that they will maintain participant confidentiality and/or anonymity. Even though Amazon prohibits Requesters from requesting personally identifiable information and MTurk Workers benefit from several features designed to protect their identities (Amazon Mechanical Turk, n.d.-c), such as anonymized Worker IDs, instances might occur where a Worker reveals their identity to the researcher.

Though unlikely to affect most researchers, Requesters should also be aware of the potential tax implications of data collection on MTurk. If a Requester pays an individual Worker more than $600 in a fiscal year, the U.S. Internal Revenue Service (IRS) requires the Requester to send the Worker a 1099-MISC form for tax purposes. When necessary, Amazon will provide Requesters with Worker information, such as name, Social Security number, and address to satisfy this requirement. The potential for Requestors to receive such sensitive personal information only increases the importance of maintaining Worker confidentiality.

A more likely disclosure occurs when an MTurk Worker communicates with a Requester via email (Mason & Suri, 2012). Workers will often use personal email accounts to ask questions, raise concerns, or dispute the rejection of submitted work. Such communication is highly likely to include the Worker's MTurk ID. Hence, researchers must take their responsibility to protect participant information seriously and prevent any knowledge of identifiable participants from influencing how they conduct or analyze the data from the study.

Jia et al. (2017) recommend the collection of IP addresses when conducting external HITs. We disagree with this practice for multiple reasons. First, collecting IP addresses has the potential to identify Workers. Second, the use of proxies, such as a virtual private network (VPN) or the Tor anonymity network, reduces the reliability of IP addresses as an indicator of a user's physical location. Third, legitimate Workers might share the same IP address, which further decreases its usefulness in screening participants. Jia et al. (2017) also suggest including additional qualitative questions with the intent to establish Worker identity while simultaneously mentioning in a footnote that requesting or collecting personally identifiable information is against Amazon's terms of service. Not only does this violate Amazon's policies, it is also likely to violate IRB guidelines. Any effort to reduce Worker anonymity is unethical and will likely result in negative Requester reputation for the researcher since review sites now prompt Workers to report violations of MTurk's terms of service (see Figure 1). Additionally, the anonymity that MTurk provides Workers should be viewed as an advantage because anonymous participants are less likely to succumb to social desirability bias. Therefore, we highly discourage researchers from collecting or requesting any such information.

4. Best Practices Before Collection

Regardless of the participant recruitment method employed, researchers must carefully plan their studies. However, this is especially true for data collected on MTurk due to the freedom and flexibility it provides. The best practices suggested in this section, summarized in Table 4, help researchers establish proper methods for soliciting, identifying, and collecting high-quality respondents from the desired population when using MTurk.

Number	Best Practice	How to Implement
4.1	Create & Secure Amazon Accounts	• Create accounts for MTurk and AWS • Enable two-step verification • Adopt a generic Requester name • Use unique email addresses for each account
4.2	Create a Qualification Test	• Create custom MTurk Qualification Types • Ask study-specific qualification questions • Broadly state HIT title, instructions, and qualification test items • Set the HIT visibility to private
4.3	Filter Workers	• Restrict access using MTurk features • Require no greater than a 97 percent HIT approval rate • Consider limiting the number of HITs approved
4.4	Generate Unique Completion Codes	• Randomly generate and assign a unique completion code to each respondent
4.5	Test Your HITs	• Use the MTurk Developer Sandbox • Collect a pilot batch before the full collection • Include qualitative questions to identify issues

Table 4. Best Practices Before Collection

4.1. Create & Secure Accounts

We suspect that many researchers new to MTurk already have personal Amazon.com accounts they use to purchase goods and services online. Nevertheless, we recommend that researchers create separate accounts on Amazon when conducting research on MTurk for a few reasons. First, proper account security includes using unique logins for each account. If one account is compromised, access to additional accounts will not be affected. Given the sensitive nature of academic research and the assurances of anonymity and confidentiality given to participants, researchers should also enable two-step verification, which is available in the Advanced Security Settings under Login & Security. Second, we advise that researchers adopt a generic Requester name. Using a personal Amazon account on MTurk can result in the researcher's name being revealed as the Requester for each HIT. While we do encourage researchers to share identifiable contact information with participants, disclosing such information can be done within the HIT instructions rather than Requester name. Third, when creating separate Amazon accounts for research purposes, we suggest that researchers use unique, non-work email addresses. If/when a researcher changes employer, they run the risk of losing access to their Requester account since Amazon's only method of contact and verification for MTurk Requesters is the email address associated with the account. This would be especially unfortunate for researchers with positive Requester reputation ratings.

While a standard Amazon.com account is all that is needed to access MTurk, we also recommend creating an Amazon Web Services (AWS) account (https://aws.amazon.com) to access the advanced features available through MTurk Developer Tools (https://requester.mturk.com/developer). Leveraging the capabilities of the MTurk Developer Tools and the associated Application Programming Interface (API) allows researchers to create Qualification Tests and to test the functionality of their HITs prior to collecting data. Amazon provides a helpful chart, which has been reproduced in Table 5 and outlines the other major benefits of using the command line tools and API as opposed to the standard web interface. The Developer Tools require that the Requester’s account be linked with an AWS account. Doing so allows the Requester to register for the MTurk Developer Sandbox and download the AWS Software Development Kit (SDK).

Creating and Managing Your Work	**Web Interface**	**Command Line Tools**	**API**
Start with our sample HTML templates	✓		
Create HITs visually with an HTML editor	✓		
Create and manage your HITs in batches	✓	✓	
Manage HITs created via the CLT or API		✓	✓
Define HITs in XML		✓	✓
Host HITs on your own server		✓	✓
Can be integrated into back-end systems			✓
Create notifications indicating when HITs are updated			✓
Managing the Workforce	**Web Interface**	**Command Line Tools**	**API**
View Worker Approval Rate on your HITs	✓		
Create custom Qualifications	✓	✓	✓
Assign a Worker a Qualification	✓	✓	✓
Revoke a Worker's Qualification	✓	✓	✓
Use system Qualifications with your HITs	up to 5	up to 10	up to 10
Use custom Qualifications	up to 5	up to 10	up to 10
Block Worker from submitting future HITs	✓	✓	✓
Remove a block from a Worker	✓	✓	✓
Give a Worker a bonus	✓	✓	✓
Email a Worker			✓

Table 5. Tool Comparison Table (reproduced from Amazon Mechanical Turk, 2018)

4.2. Create a Qualification Test

There are a few different approaches to qualifying Workers for HITs. Requesters can create a separate HIT for a qualification survey, include qualification questions at the beginning of a study, or use custom Qualification Types. We do not recommend creating a separate HIT for qualifying participants unless the study’s budget allows for offering a higher than usual payment for a longer qualification survey. Workers tend to set alerts and sort the available HITs by the reward amount. Since most qualification surveys are likely to be short and low paying, these HITs will be buried at the bottom of the list and result in a slower qualification process. We also do not recommend including qualification questions in the research instrument itself. This is likely to frustrate Workers who fail to meet the desired qualifications after beginning a HIT since it might be perceived as a “bait-and-switch” tactic.

Instead, we recommend the use of custom MTurk Qualification Types. This allows Requesters to limit the availability of a costlier and time-consuming HIT to only those who meet the desired criteria. This can be achieved prior to full-scale data collection by limiting the HIT only to Workers who have successfully obtained a custom Qualification Type. One of the major advantages of custom Qualification Types is that there are no fees paid to Workers who attempt to qualify. Workers only earn compensation after successfully qualifying and having their work accepted for the HIT. Since a high paying HIT can attract many potential participants, using custom Qualification Types is a highly cost-effective method of qualifying participants for targeted samples. However, since Requesters do not incur fees when using the custom Qualification Type method, we encourage researchers to be mindful of the Workers’ time by limiting the length of the qualification and only including items or tasks that are necessary for determining eligibility. Abusing the Qualification Type to avoid paying Workers is highly unethical and might result in the Requester’s account being terminated by Amazon.

When developing a Qualification Test, researchers should carefully consider qualification requirements and make sure the sample characteristics are as close to the target population as possible (Cheung et al., 2017; Lowry et al., 2016). Cheung et al., (2017, pg. 357) suggest including "questions that would only be answered affirmatively by someone who had the desired characteristics, such as their job title, work schedule, and salary." We discourage aggressive attempts to verify the employment status of Workers as it would violate their anonymity. However, we agree that ability can be assessed by having participants demonstrate that they have the requisite knowledge, skills, and abilities of the desired population.

Ensuring that the HIT title, instructions, and qualification test items are broadly stated is important so that Workers are not influenced to answer dishonestly simply to meet the desired target population characteristics. Signals that would reveal to participants the purpose of the study or eligibility requirements should be avoided (Cheung et al., 2017) and neutral wording can also help alleviate social desirability bias (Jia et al., 2017). For example, if a study calls for a sample of full-time employees who hold management positions at publicly traded firms in the United States, Workers might be asked to answer: 1) Please indicate your current employment status [35 hours a week or more; Less than 35 hours a week; I am not currently employed]; 2) Which of the following most closely matches your position in the organization? [intern; entry level; manager; owner]; 3) Please indicate whether your firm is privately owned or publicly traded [privately owned; publicly traded; not applicable]; 4) Which of the following best describes the organization of your employer? [for-profit; not-for-profit; government; other]. Based upon these example survey items, one could program the Qualification Test to automatically grant the custom Qualification Type to Workers who report working 35 hours or more in a management position for a publicly traded, for-profit organization.

Researchers can use Amazon's MTurk Developer Tools to create and manage custom Qualification Types. The Quiz Qualification method allows for automatic approval of Workers that meet the specified criteria, permitting qualified Workers to participate in the study immediately. Chandler, Mueller, and Paolacci (2014) provide detailed instructions on how to assign qualifications using command line tools or the web interface. Qualification surveys can also be used to ask subjects if they would like to be contacted about future studies (Mason & Suri, 2012) in order to form a pool of Worker IDs for further research (Chandler et al., 2014). If researchers are planning to conduct multiple studies that require different target populations, the relevant qualification questions can be combined into a single survey HIT. The data can then be used to generate multiple MTurk Qualifications. However, this should be done as a standalone HIT with appropriate compensation provided.

Lastly, if a Custom Qualification Test is being used, we also suggest setting the HIT Visibility to private, which is in the Worker Requirements section of the HIT properties. This allows the HIT to be visible to all Workers, but only those who have successfully obtained the custom Qualification can preview the HIT. Those who have not yet qualified will be provided a link to the Qualification Test.

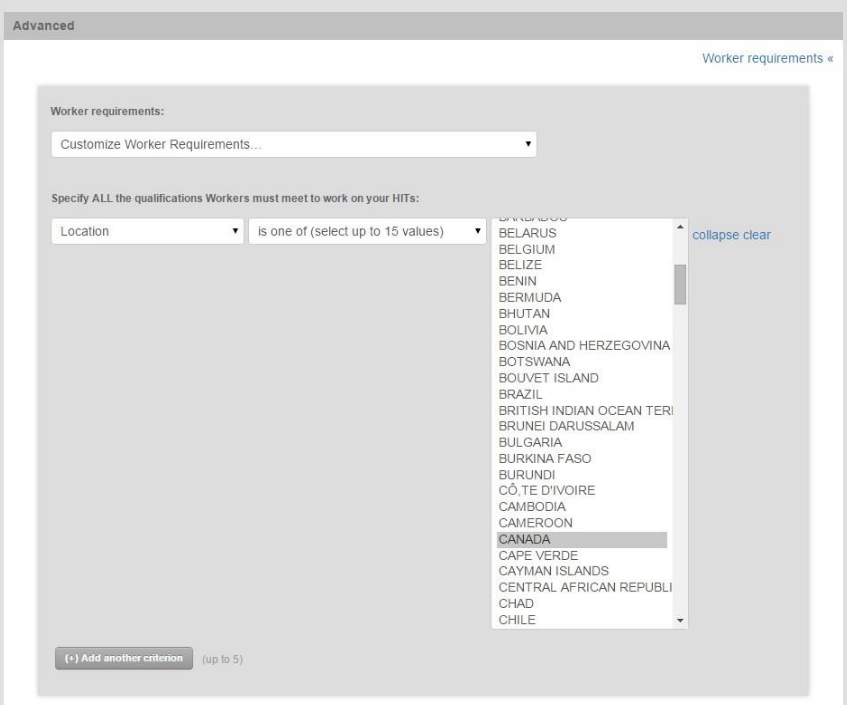
4.3. Filter Workers

Generalizability and the ability to achieve reliable statistical inference is a primary concern in academic research. One of the most critical steps before collecting data is ensuring that the methods used will result in a representative sample drawn from the target population. In addition to custom Qualification Types, MTurk also provides several features that researchers can use to filter the number of eligible respondents, such as the master qualification, premium qualifications, HIT acceptance rate, Worker location, and number of HITs accepted.

MTurk offers Requesters the ability to restrict acceptance of HITs to MTurk "Masters", who are Workers deemed by Amazon to be high-quality participants. However, there are a few drawbacks to the use of Masters. First, the process for attaining Master status is not transparent, forcing Requesters to blindly accept Amazon's judgment of Worker quality. Second, the use of Masters increases the cost of data collection by an additional five percent. Third, the pool of MTurk Masters is highly unlikely to be representative of the target populations for most research. We do not recommend using Masters unless convincing justification can be given.

MTurk introduced Premium Qualifications in 2016 (Amazon Mechanical Turk, 2016). Premium Qualifications are an attempt to categorize Workers based on regularly-sought criteria instead of forcing Requesters to include additional qualification questions into each HIT. For an additional fee, ranging from $0.05 to $1.00 per assignment, Requesters can filter the pool of Workers by the predefined list of over 130 Premium Qualifications, such as gender, industry, employment status, and job function. While the idea behind Premium Qualifications is attractive, especially to those publishing HITs that do not require accurate samples of certain populations, we discourage academic researchers from using this feature for a couple of reasons. First, just as with other online panel services, researchers cannot verify the methodology employed by MTurk. Although we recognize that the introduction of Premium Qualifications is likely to reduce the chances of Workers changing their reported characteristics from HIT to HIT, we recommend that Requesters perform their own Qualifications to maintain control and transparency. Second, since there is no additional cost for researchers to create a custom Qualification Test to assign custom Qualification Types to Workers, the surcharge for Premium Qualifications makes this feature less appealing.

Requesters can set eligibility criteria for each HIT using MTurk's built-in features for filtering participants, such as location and approval rate. Since Amazon has strengthened the standard location field for MTurk's Workers by forcing the disclosure of tax information, it can now be used reliably. This can be selected under the "Advanced" tab when setting up a HIT on MTurk by choosing to "Customize Worker Requirements", as shown in Figure 3. Requesters can then set the Worker's location as a qualification filter, allowing or restricting Workers based upon the region(s) selected.


Advanced

Worker requirements:
Customize Worker Requirements...

Specify ALL the qualifications Workers must meet to work on your HITs:
Location | is one of (select up to 15 values) | CANADA

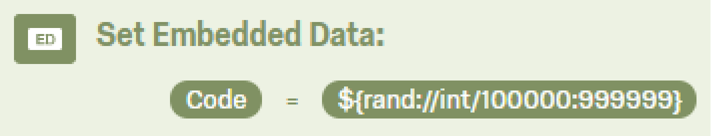
Figure 3. Location Worker Requirement

The reputation of MTurk Workers has been found to be an accurate predictor of Worker quality and successful completion of attention check questions (Peer et al., 2014). Cheung et al. (2017) and Jia et al. (2017) also mention the usefulness of Worker reputation. By using these filters, Requesters can be sure that Workers have achieved a desired Worker HIT approval rate and have had an acceptable number of HITs approved. Amazon suggests requiring Workers to have at least a 95 percent approval rate and 1,000 approved assignments (Amazon Mechanical Turk, n.d.-a). However, we feel that this recommendation is likely too restrictive for most academic research. Although most of the work that is conducted on MTurk consists of short, repetitive tasks, we suggest requiring no greater than a 97 percent HIT approval rate to allow for Workers who have up to a three percent rejection rate to participate. Since MTurk assigns a 100 percent approval rating to Workers who have completed fewer than 100 HITs, we also suggest that researchers set a minimum of 100 approved HITs to ensure that the approval rating is effective, and that Workers have some familiarity with MTurk before participating.

Jia et al. (2017) suggest increasing the sample size to lower the proportion of professional MTurk Workers. While limiting the number of professional Workers might be desirable for certain studies, simply collecting more responses is unlikely to significantly alter the proportion because the entire population of Workers has an equal opportunity to participate. Instead, if professional Workers are undesirable, we suggest that researchers consider filtering out professionals by limiting the number of HITs approved to fewer than 10,000 with an additional HIT qualification. However, researchers should be aware that using more restrictive reputation thresholds could potentially skew the participant pool and relying solely upon the HIT acceptance rate and the number of accepted HITs to qualify Workers is not advisable. Also, "professional" survey takers are not unique to MTurk as they are just as likely to participate in studies facilitated by other online panel providers, which prevents researchers from having any control over the qualification process.

4.4. Generate Unique Completion Codes

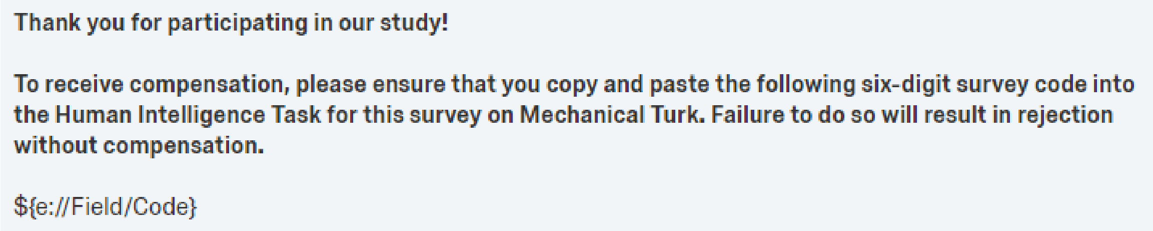
Researchers usually prefer to use other platforms that are better suited for collecting such data in conjunction with MTurk. For example, researchers can include a link in the HIT to their study instrument on Qualtrics or SurveyMonkey. Therefore, researchers need to be able to determine that a Worker claiming to have completed the HIT has in fact submitted the data collected in another platform. The most common method is to use a completion code to approve external hits (Mason & Suri, 2012), such as a combination of letters and/or numbers (e.g. “U8L4F9”) at the conclusion of the study that the Worker can enter into MTurk after participating. Some researchers might elect to use a static code for each batch to avoid verifying unique codes for each submission, although this increases the risk of participants sharing the code with other Workers to obtain payment for a HIT they did not complete. As we discuss in Best Practice 6.1, the verification process for unique completion codes can be quite painless if the researcher is comfortable with basic functions in Microsoft Excel. Therefore, we recommend the use of unique completion codes that are randomly generated and assigned to each participant. This can be achieved in Qualtrics through the built-in random number generator. For example, a random, six-digit, numeric completion code can be generated and stored in an Embedded Data field (Figure 4), and then be displayed in the survey using Piped Text (Figure 5). Though this approach would allow for the possibility of the same completion code to be assigned to multiple respondents, randomization and a large range of values makes this an unlikely event.



Set Embedded Data:

Code = ${rand://int/100000:999999}

Figure 4. Creating the Completion Code as an Embedded Data Field in Qualtrics



Thank you for participating in our study!

To receive compensation, please ensure that you copy and paste the following six-digit survey code into the Human Intelligence Task for this survey on Mechanical Turk. Failure to do so will result in rejection without compensation.

${e://Field/Code}

Figure 5. Using Piped Text to Display the Completion Code in Qualtrics

4.5. Test Your HITs

Poorly implemented HITs are likely to result in Workers leaving negative ratings and comments, so we highly recommend that researchers carefully test each aspect of the HIT, including any Qualification Tests, to confirm that they perform as intended. The first step is to test the HIT in the MTurk Developer Sandbox. Although the sandbox mimics the functionality of the production environment, HITs published in the Developer Sandbox are not visible to Workers. This allows Requesters to experiment with new uses for MTurk and to test the behavior of their HITs prior to publishing.

Once the researcher is satisfied with a HIT in the Developer Sandbox, we recommend that it be published in the production environment using a pilot batch with a limited number of assignments to verify that nothing was overlooked during the sandbox testing. Researchers should treat the pilot batch as if it were a real collection to test their rubric and approval methods. Work from the pilot batch should be classified according to the Work Approval Matrix, but we encourage Requesters to approve most work unless the Worker clearly did not put forth a reasonable effort.

Researchers should also include additional qualitative questions to identify any issues related to the technical behavior of their HIT and to gauge Worker opinion on the planned compensation relative to completion time. For example, replicating questions from the Turkopitcon review form would provide insight on how Workers are likely to view the HIT and allow researchers to mitigate these concerns prior to full data collection. Once the HIT is ready for full data collection, researchers should reduce the number of HIT-related questions as much as possible and include an open-response text box for Workers to relay any issues.

5. Best Practices During Collection

The nature of conducting studies using MTurk requires researchers to pay close attention to their HITs while they are in progress. As will be discussed in section six, more detailed data analysis should always be conducted to determine whether responses are suitable to be included in the study. In this section, we focus on the suggested practices to follow while collecting data on MTurk to ensure timely decisions can be made with respect to evaluating work on MTurk. The best practices we recommend in this section are summarized in Table 6.

Number	Best Practice	How to Implement
5.1	Capture MTurk Worker ID	• Use JavaScript to capture and store participant Worker IDs with individual responses
5.2	Repeat Study-Specific Qualification Questions	• Compare responses from qualification test to verify participant consistency and honesty
5.3	Collect Data in Batches	• Collect data in multiple batches • Resolve issues before full data collection
5.4	Promptly Remove Disqualified Participants	• Include attention check, manipulation check, and ability questions • Automatically disqualify Workers who exceed acceptable quality control thresholds • Automatically categorize Workers who are removed from the study
5.5	Exclude Repeat and Ineligible Participants	• Prevent repeat responses by excluding Worker IDs collected from prior attempts • Employ multiple approaches when excluding Workers

Table 6. Best Practices During Collection

5.1. Capture MTurk Worker ID

One of the most useful pieces of information that a researcher can gather while conducting studies on MTurk is the Worker ID, which is a randomly generated string of thirteen or fourteen alphanumeric characters assigned by Amazon to each MTurk Worker account. The Worker ID can be used to establish MTurk Qualifications, ensure the same respondent is participating in longitudinal studies, or to exclude past participants from repeated attempts. However, the Worker ID is not associated with data collected outside of MTurk without following additional steps. Some researchers might simply ask Workers to enter their Worker ID in a field within the study. However, this approach is likely to result in errors, especially if the Worker mistakes certain characters for numbers and vice versa. Even copying and pasting Worker IDs might result in extra spaces being appended to the end. Both issues can complicate the work approval process when an exact match for a given Worker ID cannot be found, potentially resulting in erroneously rejecting otherwise acceptable work. Therefore, we recommend that researchers use a script to append the Worker ID to the end of the URL for the study to automatically associate it with the participant's response on the researcher's platform of choice. A straightforward set of instructions for obtaining the Worker ID from MTurk and collecting it in Qualtrics was provided by Peer, Paolacci, Chandler, & Mueller (2012) and was later extended by Shawn Zamechek (2015). Since the Worker ID serves multiple purposes not available with other methods, we highly recommend that researchers take advantage of this feature and make certain that the Worker ID is accurately captured for each response. We have provided a modified version of this code as a supplementary file. The result of the code can be seen in Appendix C.

5.2. Repeat Study-Specific Qualification Questions

If MTurk Qualifications have been established prior to collecting study data, we suggest that the full-scale research instrument repeat the same qualification questions to verify their accuracy. By comparing the answers for each respondent from the qualification survey and the full-scale data collection, researchers can identify questionable participants. This helps eliminate any Workers who might have answered dishonestly or simply guessed the desired target population characteristics during the qualification survey, as well as Workers who might have experienced a change in their demographic status (e.g., changed jobs) between answering the qualification survey and participating in the full study.

5.3. Collect Data in Batches

Although the number of eligible participants available on MTurk is dependent upon the target population, we advise researchers to collect data in multiple batches due to the speed in which HITs are attempted by Workers. Issues that might arise during data collection are difficult to address while hundreds of attempts are in progress. Limiting the size of each batch can avoid this problem (Mason & Suri, 2012). This is especially important for researchers with limited budgets since it would be unethical to withhold payment due to any unforeseen issues with the data collection. Therefore, starting with a smaller test batch is encouraged before collecting larger sample sizes.

Another reason for employing batch collection is the short time between the initiation and conclusion of a HIT. Even though it is possible to collect thousands of responses quickly, there could be unknown issues with generalizability due to temporal bias if a sample is collected over such a narrow timeframe (Casey, Chandler, Levine, Proctor, & Strolovitch, 2018). Thus, it would be advisable to collect data in smaller batches that are initiated at different times and days of the week. One should also keep the target population in mind when developing a collection schedule. For example, unless tax season is particularly relevant to a study’s purpose, it would not be wise to seek participation from tax preparers in the United States during late March or early April because it is unlikely for the true target population to be active and fully represented on MTurk while experiencing an increased workload.

Lastly, Amazon changed the cost structure of MTurk in 2015. Previously, the fee charged to Requesters for conducting work on MTurk was 10 percent of the amount paid to Workers, including bonus payments. However, the fee is now 20 percent for HITs with up to nine assignments and 40 percent for HITs with 10 or more assignments. While the additional work involved in manually managing nine assignment HITs would be considerable, conducting small batches is a way for Requesters to reduce the cost of conducting a study on MTurk. Fortunately, the batch creation process can be automated using various programming languages, such as Python (“MTurk Documentation for Boto 3”, n.d.) and R (Carter, 2017), which helps reduce cost and avoid temporal bias.

5.4. Promptly Remove Disqualified Participants

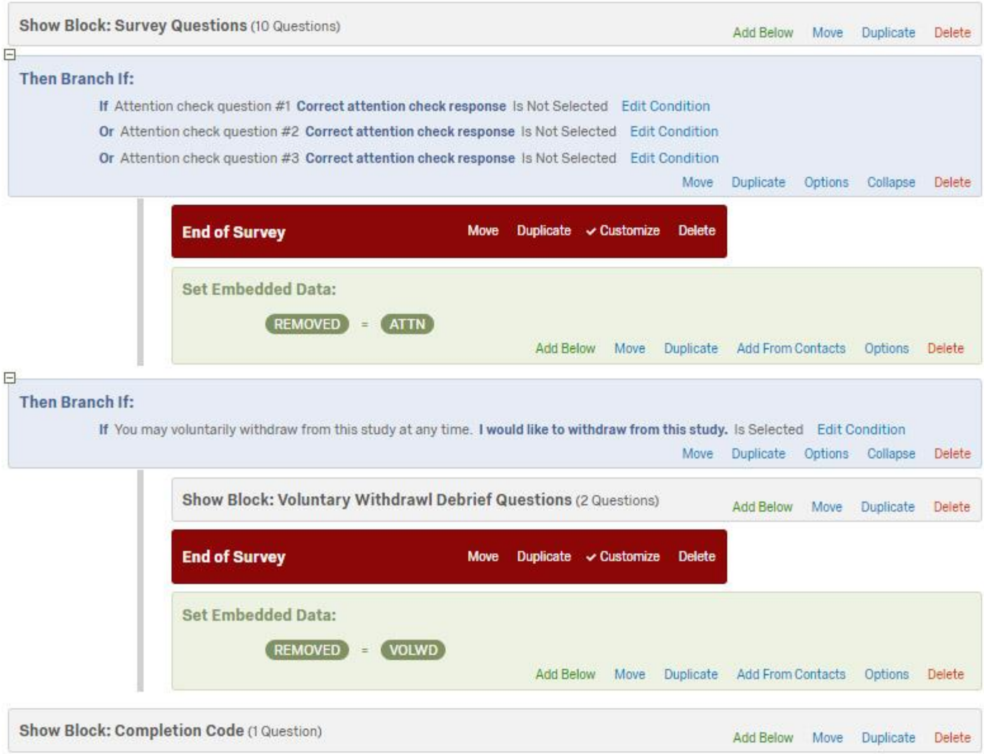
Common techniques for ensuring data quality in academic research include the use of attention check questions (ACQs), reverse-coded questions, and manipulation check questions (MCQs) (Cheung et al., 2017; Jia et al., 2017; Lowry et al., 2016; Oppenheimer, Meyvis, & Davidenko, 2009). Mason & Suri (2012) also encourage including questions that discourage spammers and bots. Lastly, researchers should avoid questions with answers that are easily found online (Goodman et al., 2013).

Establishing criteria and methods for assessing data quality should be done for all research studies, but the use of MTurk introduces unique issues that researchers must consider when identifying and removing disqualified participants. Jia et al. (2017) recommend removing participants who fail quality controls after data collection. Some have even suggested that allowing participants to have multiple attempts to complete the study would improve data quality (Cheung et al., 2017). Sprouse (2011) suggests increasing the desired sample size by 15 percent to account for rejection rates. However, we disagree with these approaches. First, we argue that researchers should set *a priori* thresholds for what is an unacceptable number of failed checks for a given study. Second, researchers should use survey logic to promptly remove participants who have exceeded quality control thresholds and prevent them from reattempting the study. Third, if you’re following Best Practice 5.3, you can simply collect additional batches until the desired sample size has been obtained. Following these recommendations will prevent the final sample from including data from inattentive participants and avoid researchers unnecessarily paying for additional attempts that should not be kept in the final sample.

Jia et al. (2017) also note that some IRBs might feel that disqualifying and removing participants violates their right to withdraw from a study without penalty or loss of benefit. However, we argue that being disqualified from a study for inattentiveness is not equivalent to voluntarily withdrawing from a study. Further, we feel that this situation can be avoided by including a “withdraw from the study” option on all screens of the study instrument. This allows participants to voluntarily remove themselves from the study and be directed to a short survey on why they wish to withdraw. Not only does this allow researchers to be notified of any concerns as they occur, but it also allows for such instances to be handled on a case-by-case basis. Researchers can still provide reasonable compensation to Workers through the bonus payment feature on MTurk.

Failing to compensate and/or communicate with disqualified participants is likely to result in reduced reputation ratings for the Requester. Therefore, consistent with the earlier recommendation to clearly communicate with Workers, we encourage researchers to include notification messages if a Worker's participation is terminated for any reason. This should be incorporated into the study design to inform Workers of the general reason for their removal (i.e., "Your responses failed to meet our quality control standards"). If the notification message is too specific (i.e., "You answered attention check questions incorrectly"), the disqualified Worker can compromise the study's integrity by warning potential participants. Informing these participants that they will still receive compensation for the time spent working on the HIT should also be included in disqualification messages when appropriate.

Manually determining the proper payment for these participants is more laborious because their work will not appear on MTurk since they are unable to submit a completion code. However, if additional embedded data fields are associated with each notification message, researchers can quickly analyze the entire data set to identify those who were disqualified for various reasons. In Qualtrics, this can be achieved using branch logic in the Survey Flow. A standard field can be set using the "Flag Response As Screened-Out" option in a custom end of survey message. However, we recommend using a custom embedded data field so that multiple values can be stored that indicate when and why the participant was removed from the study, as shown in Figure 6. In this example, the first if statement will be triggered if a participant incorrectly answers one of three attention check questions. The participant will be immediately removed from the study, provided a custom end of survey message, and the REMOVED embedded data field associated with their response will show ATTN. The second if statement will be triggered if the participant elects to voluntarily withdraw from the study. Before receiving this custom end of survey message, they will be redirected to additional questions to solicit feedback on why they elected to withdraw. Incorporating automated categorization logic simplifies the review process, especially when hundreds of responses are being collected. Simply reviewing the REMOVED embedded data field allows researchers to quickly pay these participants by uploading a batch of Worker IDs to be awarded a bonus payment. Even though the data from disqualified participants is likely unusable, clear communication and reasonable compensation is still greatly appreciated by Workers and helps encourage future participation in behavioral research.



Show Block: Survey Questions (10 Questions)

Then Branch If:

If Attention check question #1 **Correct attention check response** Is Not Selected

Or Attention check question #2 **Correct attention check response** Is Not Selected

Or Attention check question #3 **Correct attention check response** Is Not Selected

End of Survey

Set Embedded Data:

REMOVED = ATTN

Then Branch If:

If You may voluntarily withdraw from this study at any time. **I would like to withdraw from this study.** Is Selected

Show Block: Voluntary Withdrawl Debrief Questions (2 Questions)

End of Survey

Set Embedded Data:

REMOVED = VOLWD

Show Block: Completion Code (1 Question)

Figure 6. Example of Flow Logic in Qualtrics for Removing Participants

5.5. Exclude Repeat and Ineligible Participants

If researchers use MTurk to recruit their sample but collect the study data outside of MTurk, a Worker can accept a HIT, partially complete the study instrument, and then attempt to restart using the same link. Even if researchers inform

Workers that subsequent attempts to complete the HIT will not be accepted, a determined few will likely still try to participate. Since MTurk can only prevent Workers from accepting a HIT more than once, the researcher might find that they have multiple attempts from the same Worker ID despite only seeing them accept the HIT once. Most online data collection platforms attempt to prevent "ballot stuffing" (e.g., restricting participation to one response per IP address) (Lowry et al., 2016). However, these measures are not always reliable, especially if participants employ proxies, use multiple Internet browsers, or clear browser cookies.

Also, if researchers follow the suggestion to collect data in batches, previous participants must be prevented from participating in the same study again (Cheung et al., 2017). Workers who failed attention or manipulation checks in a prior attempt should be excluded from future batches of the same HIT. Failing to remove inattentive participants or allowing multiple attempts would likely invalidate their responses due to priming. Since they have been previously exposed to the HIT, researchers are no longer able to capture their true and unbiased response. This issue is especially critical for those conducting experiments. Once a Worker has been exposed to a treatment, the study would suffer from poor experimental control if he or she is given an opportunity to participate in the study a second time. Paolacci et al. (2010) also recommend tracking participants to ensure independent responses when publishing multiple HITs for the same or related studies. Additional steps must be taken to ensure that each observation collected on MTurk is unique and unbiased.

Researchers should employ multiple approaches when excluding Workers. First, the HIT should be limited to one attempt per Worker. Second, researchers should employ the method suggested by Peer et al. (2012) and use JavaScript to append the MTurk Worker ID to the link to the study instrument, which is incorporated into the code we provided in a supplementary file. Automatically checking a Worker ID against a list of previous participants is highly effective for excluding Workers when using Qualtrics. A similar approach can be adapted to other platforms using Unique Turker (http://uniqueturker.myleott.com). Lastly, researchers can completely block Workers from a HIT using MTurk's web interface or command line tools (Cheung et al., 2017).

6. Best Practices After Collection

The practices suggested in this section, summarized in Table 7, assist researchers in assessing the quality of the data collected. Researchers should perform the steps in best practices 6.1 and 6.2 immediately following the completion of each batch of data collected so that the work approval process can be completed in a timely manner.

Number	Best Practice	How to Implement
6.1	Promptly Review Submitted Work	• Check for repeat attempts based on Worker ID • Verify completion codes match each Worker ID • Review completion times for outliers • Evaluate and approve work using predetermined rubric in accordance with approval matrix • Automate steps using MTurk Developer Tools
6.2	Backup and Secure Data	• Backup all MTurk data and study responses • Disassociate Worker IDs from responses after work has been reviewed
6.3	Assess Overall Data Quality	• Check reverse coded items • Assess participant drop-out rates across treatments of an online experiment • Look for patterns in the responses • Ensure data is representative of population

Table 7. Best Practices After Collection

6.1. Promptly Review Submitted Work

We suggest researchers actively monitor and review submissions as they are completed. This allows for speedy approval and rejection and ensures that poor work is not automatically approved once the time set for auto-approval has expired. Prompt approval of work is well received by the Worker community and will be reflected in the Turkopticon reviews for the Requester, further improving the researcher's reputation. While more detailed data analysis should be reserved until the full data collection has ended, we recommend the following steps be completed in sequential order following each batch.

First, assuming that priming is a concern for the study, we recommend that researchers double check their study responses for duplicate Worker IDs. Again, best practices 5.5 and 6.1 will only prevent known Worker IDs from reattempting the study. Therefore, it is possible for Workers to accept a HIT and access the external instrument multiple times during the same batch. If multiples of the same Worker ID are present, we recommend rejecting all work associated with the Worker ID and flagging their responses to be removed from the final analysis. The prohibition of reattempts should be made clear in the study expectations (see Appendix A).

Second, we recommend that researchers check the randomized completion codes to ensure that the correct Worker ID is associated with a single, complete response. Incorrect completion codes are grounds for rejection. Completion codes entered on MTurk can be quickly matched with the Worker IDs associated with each response using a spreadsheet application, like Microsoft Excel. There are multiple functions available in Excel to assist in completing this step, such as VLOOKUP, MATCH, or INDEX.

Third, we recommend evaluating complete responses in accordance with the quality standards developed when following Best Practice 3.4. One of the most telling metrics for data quality, especially when collecting online data, is the completion time per observation (Lowry et al., 2016). If the study instrument is delivered using Qualtrics, timing questions can be embedded in each page to provide even more detail (Qualtrics, n.d.). Some respondents might be exceptionally quick readers, but the unsupervised nature of online sampling does allow for unrealistic completion times. The use of attention and manipulation check questions should catch a large majority of participants who are not reading carefully and fail to provide thoughtful responses, but a review of extreme outliers with unrealistic completion times is always a good practice. However, the decision to reject such data is far more challenging, especially if the participant successfully navigated through the attention and manipulation checks. In these select cases approving the work is probably best, but researchers might consider marking the observations as potential candidates for removal during the final data analysis. It would also be helpful to the research community, but certainly not expected, if researchers would take the time to message such Workers to encourage them to slow down when participating in future studies.

Once the quality of each response has been categorized, researchers should follow the work approval matrix from Table 3 when deciding to accept work, issue bonus payments, and communicate with Workers. The execution of this step can be automated if researchers take advantage of the MTurk Developer Tools, as discussed in best practice 4.1.

6.2. Backup and Secure Data

Researchers should be aware that MTurk data will only be available for 120 days after collection. Because of this, we encourage researchers to immediately download and backup their qualification test data and batch results from MTurk. Researchers should also save a copy of their HIT properties and content for future reference or reuse and made available to reviewers upon request. If responses are collected using an external platform, such as Qualtrics, we advise creating a backup of that data as well.

Although the collection of personally identifiable information on MTurk is prohibited by Amazon’s terms of service, we encourage researchers to treat the responses of their participants with the utmost care. While general demographic information about each Worker can be retained to qualify participants for future studies, there is no need to store the Worker ID with their individualized responses. Therefore, once researchers have completed their review of work and processed payments, it would be prudent to disassociate the Worker ID from their submission. Doing so protects participants should Amazon’s user data ever be breached, or the Worker ID is ever found to be identifiable, as was the case in the early days of MTurk (Lease et al., 2013).

6.3. Assess Overall Data Quality

While following the suggested best practices provided in this paper is likely to produce a higher level of data quality when using MTurk, no amount of vigilance can eliminate the need for additional analysis. The chances are that some of the accepted work, upon closer examination, will not be suitable for inclusion in the final analysis. Employing traditional statistical and experimental controls should not be overlooked (Kerlinger & Lee, 2000; Pedhazur & Schmelkin, 1991; Shadish et al., 2002). Researchers should still perform commonly accepted assessments for checking the quality of data (Lowry et al., 2016). For example, researchers should still check any reverse coded items and assess participant drop-out rates across treatments of an online experiment (Rand, 2012). Researchers should also look for patterns in the answer choices that possibly indicate poor quality responses (Mason & Suri, 2012) and use known population demographics (e.g., census data) or other demographic information from prior research that draws from similar populations to make sure that the data collected is representative of the target population (Cheung et al., 2017). Further, if a qualification survey was conducted to establish MTurk Qualifications, it would be wise to compare the demographics reported in both samples for each respondent to be sure consistent and reliable responses were obtained. This will help verify that the participant recruitment methods employed did, in fact, yield the desired sample.

7. Discussion

Although MTurk can be a quick, convenient, and cost-effective, yet powerful data collection method for academic research, authors often receive negative feedback from reviewers and editors about the quality of such data. Additional scrutiny is warranted if proper measures were not taken to ensure data quality, although common criticisms often have nothing to do with the actual methods employed but rather with the use of MTurk in general. Therefore, we attempt to address these concerns in the following sections.

7.1. Recommendations for Authors

Authors should adopt as many of the suggested best practices as possible to improve the quality of data collected on MTurk. Although page limits often require authors to shorten or remove insightful explanations of the data collection procedures, we believe that providing this information is extremely valuable to assessing data quality and should, therefore, be included. Following the practices outlined in this paper would also allow authors to simply provide a citation to concisely communicate the data collection methods employed. However, authors are still encouraged to explain study-specific criteria, such as qualification questions, to provide additional insight on the methods employed to sample the desired target population.

7.2. Recommendations for Reviewers

Regardless of the platform used, reviewers should require that authors disclose their data collection procedures to better assess data quality rather than making an assessment based solely upon the platform being used. In fact, we argue that the use of MTurk affords researchers greater control and understanding of the data collection process, especially when compared to paid online panel providers that promise to deliver samples of the desired populations yet fail to provide any real method of verification. Therefore, reviewers should carefully critique the methods used for participant recruitment, qualification, and compensation for all research. It should also be noted that, unlike MTurk, the amount paid to online research companies (e.g., Qualtrics and SurveyMonkey) for online panels is not directly paid to those who participate. Considering such a rate as participant compensation or gauging the perceived "quality" of data collected based upon such a figure is inaccurate.

7.3. Recommendations for Editors

Poor practices can certainly lead to poor data, but MTurk provides researchers far more control and insight into their sample than paying other firms to recruit participants for their study. Editors should be sure that any issues raised by reviewers pertaining to the use of MTurk are based upon the methods employed by the authors rather than MTurk in general. Encouraging reviewers to critique participant recruitment, qualification, and compensation, rather than simply disregarding MTurk as a research tool, will yield constructive feedback and improve the quality of all research. While we understand the difficulty of staying under page limitations, we encourage editors to request that a detailed description of the sampling methodology be reported for every study to improve the assessment of data quality for all published research.

8. Conclusion

Although we only focused our paper on survey and experimental research, the wide range of research applications for MTurk is exciting. Regardless of how MTurk is used, researchers must make sure that proper measures are taken to maintain academic rigor. Researchers might find themselves overwhelmed with having total control of the subject recruitment and qualification process, so we have provided a practical tutorial to follow before, during, and after conducting research using MTurk. We discussed specific options and settings available in MTurk as well as included images, websites, and scripts so that researchers new to MTurk will be able to successfully create their own HITs. Following our recommended best practices should ease the burden of using MTurk and ultimately enhance the quality of data collected. We also provide arguments for the acceptance of MTurk as a quality research platform and discuss significant advantages of MTurk over existing online methods currently accepted. Finally, we argue that reviewers and editors of academic research must ensure that criticism of the data collection methods employed in any study is rooted in the procedures followed, not the platform itself.

References

Amazon Mechanical Turk. (n.d.-a). Amazon Mechanical Turk Concepts. Retrieved from https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/mechanical-turk-concepts.html

Amazon Mechanical Turk. (n.d.-b). Amazon Mechanical Turk Pricing. Retrieved July 27, 2018, from https://requester.mturk.com/pricing

Amazon Mechanical Turk. (n.d.-c). Amazon Mechanical Turk Worker FAQs. Retrieved June 10, 2018, from https://www.mturk.com/worker/help

Amazon Mechanical Turk. (2016). Introducing Premium Qualifications. Retrieved July 27, 2018, from https://blog.mturk.com/introducing-premium-qualifications-1e473456e7b0

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. https://doi.org/10.3758/s13428-011-0081-0

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. https://doi.org/10.1093/pan/mpr057

Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples. *Behavior Research Methods*, *49*(1), 320–334. https://doi.org/10.3758/s13428-016-0710-8

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An Evaluation of Amazon’s Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science*, *13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Carter, W. (2017). *Automating MTurk HIT Creation: A Manual for Overcoming MTurk HIT Price Overhauls with R*. Retrieved from http://wtcarter.web.unc.edu/files/2018/01/MTurk-Manual.pdf

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2018). *Intertemporal Differences Among MTurk Worker Demographics*. Retrieved from https://psyarxiv.com/8352x?file=57dc09c6b83f6901e2f106e3

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

Chandler, J., Paolacci, G., & Mueller, P. (2013). Risks and Rewards of Crowdsourcing Marketplaces. In P. Micheluccui (Ed.), *Handbook of Human Computation* (pp. 377–392). New York, NY: Springer Science+Business Media. https://doi.org/10.1007/978-1-4614-8806-4_30

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

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Deng, X. (Nancy), & Joshi, K. D. (2016). Why Individuals Participate in Micro-task Crowdsourcing Work Environment: Revealing Crowdworkers’ Perceptions. *Journal of the Association for Information Systems*, *17*(10), 648–673.

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., & Cudré-Mauroux, P. (2015). The Dynamics of Micro-Task Crowdsourcing. In *Proceedings of the 24th International Conference on World Wide Web - WWW ’15* (pp. 238–247). New York, New York, USA: ACM Press. https://doi.org/10.1145/2736277.2741685

Difallah, D. E., Filatova, E., & Ipeirotis, P. G. (2018). Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM ’18* (pp. 135–143). New York, New York, USA: ACM Press. https://doi.org/10.1145/3159652.3159661

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive Responding in MTurk and Other Online Samples. *Industrial and Organizational Psychology*, *8*(02), 196–202. https://doi.org/10.1017/iop.2015.25

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

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. https://doi.org/10.1002/bdm.1753

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk. In *SIGCHI Conference on Human Factors in Computing Systems* (pp. 1–14). New York, NY. https://doi.org/10.1145/3173574.3174023

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

Horton, J. J., & Chilton, L. (2010). The Labor Economics of Paid Crowdsourcing. *Proceedings of the 11th ACM Conference on Electronic Commerce*, (1), 209–218. https://doi.org/10.1145/1807342.1807376

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425. https://doi.org/10.1007/s10683-011-9273-9

Ipeirotis, P. G. (2010a). Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, *17*(2), 16. https://doi.org/10.1145/1869086.1869094

Ipeirotis, P. G. (2010b). *Demographics of mechanical turk* (NYU Working Paper Series No. CEDER-10-01). New York, New York. Retrieved from http://archive.nyu.edu/handle/2451/29585

Jia, R., Reich, B. H., & Jia, H. H. (2016). A commentary on: “Creating agile organizations through IT: The influence of IT service climate on IT service quality and IT agility.” *Journal of Strategic Information Systems*, *25*(3), 227–231. https://doi.org/10.1016/j.jsis.2016.06.001

Jia, R., Steelman, Z. R., Reich, B. H., & Jia, R. (2017). Using Mechanical Turk Data in IS Research: Risks, Rewards, and Recommendations. *Communications of the AIS*, *41*(1), 301–318.

Kaufmann, N., & Veit, D. (2011). More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems* (pp. 1–11). Detroit, Michigan.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of Behavioral Research* (4th ed.). Belmont, California: Wadsworth Publishing.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: report of Board of Scientific Affairs’ Advisory Group on the Conduct of Research on the Internet. *The American Psychologist*, *59*(2), 105–117. https://doi.org/10.1037/0003-066X.59.2.105

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*(02), 142–164. https://doi.org/10.1017/iop.2015.13

Lease, M., Hullman, J., Bigham, J. P., Bernstein, M., Kim, J., Lasecki, W. S., Bakhshi, S., Mitra, T., & Miller, R. C. (2013). *Mechanical Turk is Not Anonymous*. https://doi.org/10.2139/ssrn.2228728

Lowry, P. B., D’Arcy, J., Hammer, B., & Moody, G. D. (2016). “Cargo Cult” science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels. *Journal of Strategic Information Systems*, *25*(3), 232–240. https://doi.org/10.1016/j.jsis.2016.06.002

Mason, W. A., & Suri, S. (2012). Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Mason, W. A., & Watts, D. J. (2009). Financial incentives and the “performance of crowds.” *ACM SIGKDD Explorations Newsletter*, *11*(2), 100. https://doi.org/10.1145/1809400.1809422

MTurk Documentation for Boto 3. (n.d.). Retrieved July 28, 2018, from https://boto3.readthedocs.io/en/latest/reference/services/mturk.html

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

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. Retrieved from http://repub.eur.nl/pub/31983

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, Design, and Analysis: An Integrated Approach* (1st ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Peer, E., Paolacci, G., Chandler, J., & Mueller, P. (2012). Screening Participants from Previous Studies on Amazon Mechanical Turk and Qualtrics. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2100631

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

Qualtrics. (n.d.). Timing Question. Retrieved from https://www.qualtrics.com/support/survey-platform/survey-module/editing-questions/question-types-guide/advanced/timing/

Rand, D. G. (2012). The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179. https://doi.org/10.1016/j.jtbi.2011.03.004

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., & Vukovic, M. (2011). An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *Fifth International AAAI Conference on Weblogs and Social Media* (pp. 321–328).

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the Crowdworkers ? Shifting Demographics in Amazon Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (pp. 2863–2872). Atlanta, Georgia: ACM Press. https://doi.org/10.1145/1753846.1753873

Roulin, N. (2015). Don't Throw the Baby Out With the Bathwater: Comparing Data Quality of Crowdsourcing, Online Panels, and Student Samples. *Industrial and Organizational Psychology*, *8*(2), 190–196. https://doi.org/10.1017/iop.2015.24

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and Quasi-Experimental for Generalized Designs Causal Inference. *Handbook of Industrial and Organizational Psychology*, *223*, 623. https://doi.org/10.1198/jasa.2005.s22

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to Study Clinical Populations. *Clinical Psychological Science*, *1*(2), 213–220. https://doi.org/10.1177/2167702612469015

Shawn Zamechek. (2015). How to capture MTurk workers' IDs into your Qualtrics survey. Retrieved August 19, 2018, from https://research-it.wharton.upenn.edu/news/capture-mturk-workers-ids-qualtrics-survey/

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167. https://doi.org/10.3758/s13428-010-0039-7

Steelman, Z. R., Hammer, B. I., & Limayem, M. (2014). Data collection in the digital age: Innovation alternatives to student samples. *MIS Quarterly*, *38*(2), 355-A20.

Stewart, N., Ungemach, C., Harris, A. J. L., 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*(5), 479–491. https://doi.org/10.1017/CBO9781107415324.004

Appendix A: Recommended Language for HIT Expectations

In this appendix, we provide researchers with recommended language to use when communicating expectations to Workers in the HIT description on MTurk. Please note that the elements in brackets should be edited to fit the context of the study in question. Be sure to use high-level language to avoid priming participants.

Expected Time	Based upon average completion time from a pilot study, completing this HIT will take approximately **[X]** to **[Y] minutes**. The time allotted to complete this HIT is **[Y x 2] minutes**.
Compensation	The reward for accepted work is **[Recommended minimum: $7.25 x completion time for the 75th percentile from pilot study/60 minutes]** for this HIT.
Importance	This scientific study will impact [broadly stated research area]. Your attentive and honest responses are appreciated.
Environment	Prior to accepting this HIT, please ensure that you are in a distraction-free environment that is conducive to deep thought.
Acceptance	We will review work within **[X] hours**. Honest, attentive, and complete responses will be accepted.
Rejection	Your work will be rejected if it does not satisfy our quality standards. If your work is rejected, you will be compensated for your time through a bonus payment.
Communication	The researcher(s) may be contacted via email at any time. You will be provided with contact email addresses at the beginning of the study. However, please ensure that you use an email address that will not identify you and only refer to your work by providing your Mechanical Turk Worker ID.
Affirmation	By accepting this HIT, you affirm that you have read and understand the expectations of participating in this study.

Table 8. Example HIT Expectations

Appendix B: Recommended Language for Study Instrument

In this appendix, we provide researchers with recommended language to use on the study instrument. We recommend including Table 9 after your institution's IRB human consent form.

Contact Information	If you have any questions or concerns about the study, you may contact the researchers via email. However, please ensure that you use an email address that will not identify you and only refer to your work by providing your Mechanical Turk Worker ID. **[Researcher 1] [researcher1@example.edu]** **[Researcher 2] [researcher2@example.edu]**
Repeated Attempts	Be sure that you only click on the study link once. If you experience technical issues, please contact us immediately before reattempting the study. Unauthorized repeat attempts will be rejected without compensation.
Quality Controls	Your work will be rejected if it does not satisfy our quality standards. If your work is rejected, you will still be compensated for your time through a bonus payment. Reattempts will be rejected without additional payment.
Research Purposes	The data collected for this study will be used for academic research purposes. We intend to publish the results of this study in academic outlets, such as conferences and journals.
Anonymity	Your anonymity is important to us. We have made every effort to avoid the collection of any personally identifiable information. Unless you have indicated that you would like to be considered for future studies, the use of your Worker ID is strictly for HIT approval and payment purposes. However, if you inadvertently disclose personally identifiable information, we promise not to disclose your identity to any third-party.
Confidentiality	The responses you provide while participating in this study will be kept strictly confidential. Data analysis will be reported in aggregate form. Written responses will be anonymized, with no reference to your Worker ID.
Non-Disclosure	The content of this study is confidential and should not be shared with other potential participants (forums, social media, etc.). Doing so will jeopardize the integrity of the research project.
Feedback	You will have an opportunity to provide feedback at the end of the study. Please report any questions, concerns, and/or difficulties experienced. Your feedback will help us ensure that we provide a positive experience for other Workers on MTurk.
Affirmation	By continuing, you affirm that you have read and understand the instructions for this study.

Table 9. Example Study Overview

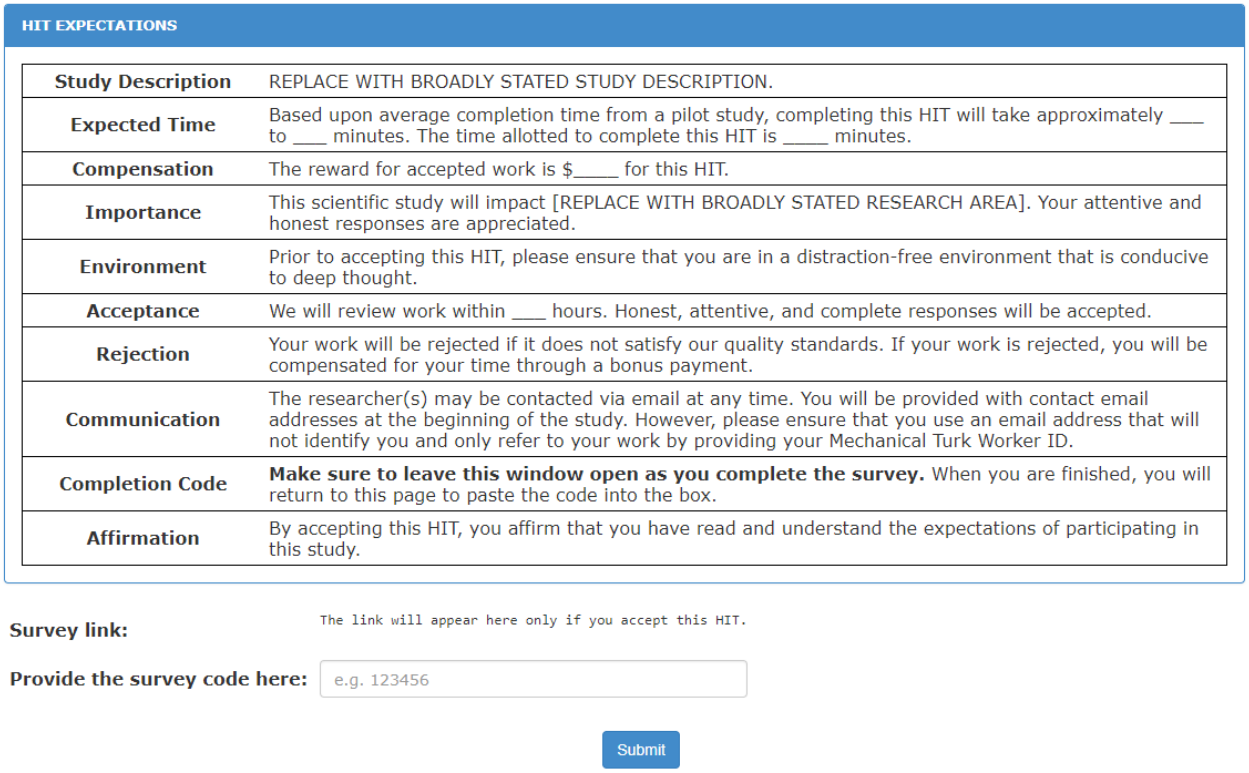
We recommend including **Table 10** as the last screen of the study instrument.

Confidentiality	Thank you for participating in our study! Remember, to preserve the integrity of the study, you may not share anything about the experiment with other potential participants (forums, social media, etc.).
Completion Code	To receive compensation, please ensure that you copy and paste the following six-digit survey code into the Human Intelligence Task for this study on Mechanical Turk. **[Randomized Completion Code]**

Table 10. Example End of Survey Screen

Appendix C: Code for HIT Expectations

Peer, Paolacci, Chandler, & Mueller (2012) provided a script to append Worker IDs to the study URL that was later extended by Shawn Zamechek (2015). We build upon their work by incorporating our suggested HIT Expectations language from Table 8 into the code we provide as a supplementary file. Replacing the default code in the Design Layout with our code will create the HIT Expectations shown in Figure 7.



HIT EXPECTATIONS

Study Description	REPLACE WITH BROADLY STATED STUDY DESCRIPTION.
Expected Time	Based upon average completion time from a pilot study, completing this HIT will take approximately ___ to ___ minutes. The time allotted to complete this HIT is _____ minutes.
Compensation	The reward for accepted work is $_____ for this HIT.
Importance	This scientific study will impact [REPLACE WITH BROADLY STATED RESEARCH AREA]. Your attentive and honest responses are appreciated.
Environment	Prior to accepting this HIT, please ensure that you are in a distraction-free environment that is conducive to deep thought.
Acceptance	We will review work within ___ hours. Honest, attentive, and complete responses will be accepted.
Rejection	Your work will be rejected if it does not satisfy our quality standards. If your work is rejected, you will be compensated for your time through a bonus payment.
Communication	The researcher(s) may be contacted via email at any time. You will be provided with contact email addresses at the beginning of the study. However, please ensure that you use an email address that will not identify you and only refer to your work by providing your Mechanical Turk Worker ID.
Completion Code	**Make sure to leave this window open as you complete the survey.** When you are finished, you will return to this page to paste the code into the box.
Affirmation	By accepting this HIT, you affirm that you have read and understand the expectations of participating in this study.

Survey link: The link will appear here only if you accept this HIT.

Provide the survey code here: e.g. 123456

Figure 7. HIT Instructions on MTurk When Using Provided Code

Author Biographies

Jacob A. Young is an assistant professor of management information systems and the director of the Center for Cybersecurity at Bradley University. He earned his doctorate from Louisiana Tech University and received his Bachelor of Science and Master of Business Administration from Henderson State University. He focuses his research on privacy, security and anonymity issues related to information systems. His work has been published in *AIS Transactions on Human-Computer Interaction* and the *DePaul Business & Commercial Law Journal*.

Kristie M. Young is an assistant professor of accounting at Illinois State University. She earned her Bachelor of Science in Accounting and Master of Business Administration from Eastern Illinois University and her doctorate at Louisiana Tech University. Dr. Young’s research interests include mergers and acquisitions and measuring the impact on psychological contracts and employee outcomes. Her work has been published in *Management Accounting Quarterly* and the *IMA Educational Case Journal*.

This page intentionally left blank