BEHAVIORAL RESEARCH IN ACCOUNTING
Vol. 30, No. 1
Spring 2018
pp. 111–122

American Accounting Association
DOI: 10.2308/bria-51977

A Technical Guide to Using Amazon's Mechanical Turk in Behavioral Accounting Research

Steve Buchheit
Marcus M. Doxey
Troy Pollard
Shane R. Stinson
The University of Alabama

ABSTRACT: Multiple social science researchers claim that online data collection, mainly via Amazon's Mechanical Turk (MTurk), has revolutionized the behavioral sciences (Gureckis et al. 2016; Litman, Robinson, and Abberbock 2017). While MTurk-based research has grown exponentially in recent years (Chandler and Shapiro 2016), reasonable concerns have been raised about online research participants' ability to proxy for traditional research participants (Chandler, Mueller, and Paolacci 2014). This paper reviews recent MTurk research and provides further guidance for recruiting samples of MTurk participants from populations of interest to behavioral accounting researchers. First, we provide guidance on the logistics of using MTurk and discuss the potential benefits offered by TurkPrime, a third-party service provider. Second, we discuss ways to overcome challenges related to targeted participant recruiting in an online environment. Finally, we offer suggestions for disclosures that authors may provide about their efforts to attract participants and analyze responses.

Keywords: Amazon Mechanical Turk; online experiments; participant screening; performance-based incentives; online labor markets; TurkPrime.

INTRODUCTION

The use of online workers, or "crowdsourcing," in academic research has increased significantly in recent years, reflecting the pace with which technology has evolved to increase the size, diversity, and overall convenience of online subject pools (e.g., Mason and Suri 2012). While this trend has "revolutionized behavioral sciences" (Gureckis et al. 2016, 829; Litman, Robinson, and Abberbock 2017, 433) and given rise to several venues for online subject recruitment, Amazon's Mechanical Turk (MTurk) platform offers several key features that improve researchers' control over participant recruiting and payment terms. These features have made MTurk a leading crowdsourcing marketplace used by academic researchers (Chandler and Shapiro 2016; Crump, McDonnell, and Gureckis 2013).[1] Supporting this notion, Chandler and Shapiro (2016, 55) document consistent and extraordinary growth in the number of published studies in psychology using MTurk data (fewer than ten papers in

We gratefully acknowledge Derek Dalton, Brian Goodson, Jonathan Grenier, Richard C. Hatfield (editor), Chad Stefaniak, workshop participants at the 2017 Behavioral Tax Symposium, and our anonymous reviewer for their thoughtful comments.

Editor's note: Commissioned by Richard C. Hatfield.

Submitted: May 2017
Accepted: November 2017
Published Online: November 2017

[1] For discussion and analysis of alternative marketplaces—which are generally comparable to MTurk—see Peer, Brandimarte, Samat, and Acquisti (2017).

2010 increasing to over 500 in 2015), and Harms and DeSimone (2015, 184) cite similar exponential growth.[2] In this article, we add to the literature on MTurk by offering technical guidance and discussing implementation challenges that may be of particular concern to behavioral researchers in accounting.

Online participants present novel research risks (e.g., participants are unsupervised and anonymous, they complete surveys in unknown locations, and they are motivated by small financial incentives) that have led to concerns about the quality of crowdsourced data (Chandler, Mueller, and Paolacci 2014; Bentley 2017). We briefly review research examining online subjects and note a general consensus that MTurk workers are largely high-quality research participants. However, to be clear, we do not argue in favor of using MTurk participants in all behavioral accounting experiments—thoughtfully matching subjects to the goals of an experiment (Libby, Bloomfield, and Nelson 2002, 802) remains sage advice. Further, the trade-offs of using online versus professional subjects are similar to the trade-offs of using students in lieu of professionals (e.g., easy access but generally low expertise), although online participant pools also have unique costs and benefits.

We consider these costs and benefits from three perspectives: instrument logistics, participant recruiting, and research disclosure. Regarding logistics, Brandon et al. (2014) discuss online instrument design for basic surveys and online delivery of experiments within accounting (e.g., using Qualtrics and SurveyMonkey). However, using MTurk can be particularly difficult for accounting experiments involving variable bonus payments (i.e., performance-based pay) and panel studies. We discuss relatively easy ways to overcome the preceding challenges and provide basic "start-up" guidance for researchers considering MTurk for the first time. For example, we discuss how to set up a user profile, pay participants, and resolve technical or communication difficulties.

Regarding participant recruiting, accounting researchers have long explored conditions in which nonprofessional participants can legitimately proxy for groups more difficult to obtain (e.g., Ashton and Kramer 1980; Elliott, Hodge, Kennedy, and Pronk 2007) and have argued that nontraditional participants are sometimes more suitable subjects than professionals because of efficiency concerns (Libby et al. 2002). In this same vein, recent accounting studies have compared MTurk participant performance to student performance (Farrell et al. 2017; Buchheit, Dalton, Pollard, and Stinson 2017) and found that MTurk participants often provide high-quality data. Nevertheless, attracting participants with particular traits or skills (e.g., taxpayers or potential jurors) is a challenge using the MTurk platform. We discuss various procedures that facilitate attracting desired MTurk participants (e.g., "invitation-only" experiments), particularly from populations of interest to accounting researchers. Finally, regarding research disclosure, we discuss the challenges of communicating the steps taken to promote data integrity at each stage of collection (e.g., initial recruiting and excluding invalid responses) to readers, reviewers, and editors.

In total, we offer guidance that addresses many common concerns and obstacles in using MTurk to recruit participants for behavioral accounting studies. We seek to aid behavioral accounting researchers in harnessing the power and mitigating potential dangers unique both to the MTurk platform and to crowdsourcing in general. The paper continues as follows. In the second section, we briefly review prior MTurk research. In the third section, we discuss the mechanics of conducting MTurk experiments. In the fourth section, we address participant sampling and reporting issues. The last section concludes.

PRIOR RESEARCH: CONCERNS ABOUT MTURK

Perhaps the most common concern about MTurk participants is that they are an inherently flawed population. Low pay rates, negligible barriers to participation, and a relative lack of experimental control drive general concerns about MTurk participants (Kraut et al. 2004).[3] More specific concerns are that the population may be less willing to exert effort, pay close attention to protocols, or honestly respond to demographic and screening questions compared to traditional research participants (Kraut et al. 2004; Oppenheimer, Meyvis, and Davidenko 2009; Paolacci, Chandler, and Ipeirotis 2010; Chandler et al. 2014; Hauser and Schwarz 2016). A number of studies speak to, and largely attenuate, the preceding concerns about MTurk (e.g., Bentley 2017; Chandler and Shapiro 2016).

For example, compared with the general population, several studies show that MTurk participants are younger, more computer literate, and more likely to be single, but they are less likely to be homeowners and religiously affiliated (e.g., Berinsky, Huber, and

[2] We examined all articles involving human subjects experiments published from 2010 through 2016 in *Accounting Horizons; Accounting, Organizations and Society; The Accounting Review; Auditing: A Journal of Practice & Theory; Behavioral Research in Accounting; Contemporary Accounting Research; Journal of Accounting Research*; *The Journal of the American Taxation Association;* and the *Journal of Managerial Accounting Research*. While published papers using crowdsourced samples in accounting has lagged other disciplines (e.g., a "high" year of seven publications occurred in 2015), we note that a growing number of working papers in accounting use MTurk samples. In addition, a well-attended crowdsourcing session at the 2016 ABO Midyear Meeting, a requested session at the 2017 Behavioral Tax Symposium, and high-profile accounting methodological papers involving MTurk participants (Farrell, Grenier, and Leiby 2017; Brandon, Long, Loraas, Mueller-Phillips, and Vansant 2014) all suggest growing interest. Thus, relative to other disciplines, accounting's apparent "slow adoption" of crowdsourced participant samples may be an artifact of delays in the publication process and/or a proportionately smaller body of experimental research rather than any widespread resistance to the platform.

[3] Early studies also examined why people choose to participate in MTurk given the relatively low pay. For example, Ross, Irani, Silberman, Zaldivar, and Tomlinson (2010) find that some workers view their MTurk earnings as an important supplement to their primary income that can be earned on a flexible time schedule. Ipeirotis (2010) finds that workers generally view MTurk as a "fruitful way to spend free time and get some cash."

Lenz 2012; Smith, Roster, Golden, and Albaum 2016). On the other hand, these studies also show that MTurk workers are more diverse than student populations commonly used in academic research. Moreover, while Amazon claims MTurk provides access to over 500,000 workers from 190 countries (https://requester.mturk.com/tour), Stewart et al. (2015) refine this measure using capture-recapture analysis and estimate an average of 7,300 workers are available at any given time. Thus, the MTurk population appears large and able to accommodate extensive demographic targeting that may not be feasible for students or other traditional subject pools (discussed in the "Attracting MTurk Participants and Disclosing Sample Selection Techniques" section).

Further, researchers in various disciplines have validated the responses of MTurk participants for consistency and psychometric properties (e.g., reliability, convergent and divergent validity). Loosely speaking, MTurk participant responses are similar to traditional subject pool responses. For example, MTurk technology acceptance responses are comparable to general consumer behavior (Steelman, Hammer, and Limayem 2014); MTurk workers' worldview and attitudes are similar to standard consumer panel attitudes (Smith et al. 2016); MTurk participant responses to many well-known psychology tasks (e.g., the Stroop task) produce results similar to prior research (Crump et al. 2013); and MTurk participant responses to standard economic games resemble those of students (Horton, Rand, and Zeckhauser 2011). In contrast to general similarities between MTurk and alternative participant pools, one consistently identified difference is that international MTurk workers respond differently and have higher error rates relative to U.S. MTurk workers and traditional subjects (Steelman et al. 2014; Smith et al. 2016).

Given that accounting tasks can be demanding compared to typical MTurk tasks, two recent methodological studies use tasks from prior accounting literature to compare the quality of MTurk workers' decisions to those of student workers. First, Farrell et al. (2017) examine MTurk workers' honesty and effort and find that, relative to students, MTurk workers are no worse (and in the case of effort, potentially better) than student participants.[4] Second, Buchheit et al. (2017) examine MTurk workers' fluid intelligence relative to students and, like Farrell et al. (2017), find that MTurk workers are no worse than traditional student participants. Both studies use performance-based pay and vary incentives at high and low levels. Both studies also find that relatively low performance-based pay levels (i.e., around the U.S. minimum wage) sufficiently incentivize MTurk workers and higher pay levels do not improve overall performance (i.e., relatively low incentive pay levels appear to maximize effort). Collectively, recent MTurk research, including accounting-focused studies, provides evidence that MTurk offers a high-quality, low-cost participant pool for relatively demanding tasks that might otherwise be completed by traditional student participants or by nonaccounting experts (e.g., see Grenier, Reffett, Simon, and Warne [2018] for a discussion of using MTurk participants as hypothetical jurors in accounting research).

CONDUCTING RESEARCH ON MTURK

Researchers outside accounting have provided detailed descriptions for many of MTurk's features that appeal broadly to experimental researchers (Mason and Suri 2012; Chandler and Shapiro 2016; Gureckis et al. 2016; Litman et al. 2017). Rather than duplicate those efforts, we first offer a brief synopsis of MTurk's core functions, highlighting common implementation challenges that may be most relevant to accounting researchers. We then focus on more recent developments that address these challenges.

At its core, MTurk offers a common platform pairing requesters (experimental researchers, marketing firms, private enterprises, etc.) who create online Human Intelligence Tasks (HITs) with workers who complete those tasks for payment. Both requesters and workers must join MTurk by linking their user profiles to a valid Amazon account. Requesters compete for the time and attention of MTurk workers based on the nature and rewards of their HITs. HITs can take many forms (e.g., directed web searches, text analysis, image captioning), although studies in accounting often rely on advanced survey formats (e.g., Brink and Lee 2015; Brink, Eaton, Grenier, and Reffett 2017; Doxey, Hatfield, Rippy, and Peel 2017; Farrar, Kaplan, and Thorne 2018b; Farrell et al. 2017).[5]

Requesters create a HIT describing the task and specifying payment terms, the number of workers needed, and both the expected and maximum time allowed for the task. The requester commits the necessary funds to his or her MTurk account (i.e., each HIT must be prepaid) before the HIT is made available to potential MTurk workers. Workers can log in and search for live HITs by keyword, duration, compensation, etc. Workers typically apply for payment using a code provided only after finishing the HIT (e.g., a code provided at the end of a Qualtrics survey). The requester must then approve or reject applications for payment, either manually or automatically, depending on the HIT settings. As a result, MTurk handles all payments electronically.

Amazon's pricing structure scales up with increases in sample size and includes additional fees when researchers use embedded screening tools. For example, Amazon currently charges a 20 percent markup on all worker payments for HITs with

[4] Similarly, Hauser and Schwarz (2016) find that MTurk workers outperform student participants with respect to experimental attention checks.

[5] Further, while MTurk provides a survey-generating platform, the current features and graphical interface of the MTurk platform are limited compared to Qualtrics and other more established alternatives familiar to accounting researchers. For this reason, accounting researchers tend to populate HITs with links to external surveys that offer greater flexibility and functionality suitable for experimental accounting tasks.

fewer than ten workers. For HITs with ten or more workers, the markup increases to 40 percent of total worker payments.[6] Extra charges apply for workers with a particular expertise or demographic trait, which Amazon refers to as "premium qualifications." However, unlike some other popular online recruitment platforms (e.g., Qualtrics Panel), researchers using MTurk control participant screening and compensation terms directly. This includes the ability to determine participant compensation levels and variable compensation, such as performance-based incentives.

Finally, while a HIT is in progress, it is not unusual for workers to have questions (due to technical difficulties or perceived instructional ambiguity), and workers often seek accommodations from the requester (e.g., Lease et al. 2013). Such questions and requests from MTurk workers are directed to the email address linked to the requester's account.[7] As described below, researchers can use third-party services, such as TurkPrime, to resolve many common technical difficulties associated with MTurk.

Limitations of MTurk and Benefits of TurkPrime

MTurk's popularity is largely due to convenient access to reasonably high-quality individuals willing to work at a low cost. However, MTurk's interface has several shortcomings that may impede data collection for more complex tasks. Fortunately, third-party intermediary services mitigate many of MTurk's limitations. Below, we discuss the advantages of TurkPrime, a third-party service that links to a requester's MTurk account.[8]

Relative to MTurk, TurkPrime uses a more intuitive graphical interface and offers several useful tools for interacting with workers and customizing HITs that are unavailable through MTurk's native interface.[9] For instance, TurkPrime affords the ability to edit HITs after they launch. This gives the researcher greater flexibility to fine-tune a study's payment structure if workers show too little interest in a HIT or change the estimated time requirements if workers cannot complete a HIT in the allotted time. TurkPrime also features a restart option that makes a new copy of an existing HIT (with automatic safeguards preventing repeated participation from the same workers) and refreshes the HIT's presence in MTurk's task listings. This can prove critical to large sample studies, as potential workers often target "fresh" HITs that are less than 24 hours old (Litman et al. 2017; Chilton, Horton, Miller, and Azenkot 2010).

In addition, while MTurk does not easily accommodate HIT invitations for specific users, TurkPrime allows requesters to specifically include individual workers and notify them of HITs with custom email invitations. Similarly, TurkPrime allows requesters to exclude individual workers from viewing a HIT. Thus, TurkPrime provides more control over sample entrants without the added steps of creating custom qualifications in MTurk, waiting for desired participants to discover a HIT, and perhaps blocking or rejecting undesired workers who try to complete a HIT that was not intended for them.[10] Further, as noted by Litman et al. (2017), the ability to limit workers' access to a HIT based on their prior participation greatly reduces the administrative challenges of conducting multi-part studies—with TurkPrime, a requester can simply run the first in a series of HITs, and then target only the previously approved workers for all subsequent HITs. This TurkPrime feature is particularly useful for longitudinal sample selection (discussed in the next section).

---

[6] Despite this seemingly aggressive markup structure, MTurk remains inexpensive relative to other online platforms. For example, paying MTurk participants $2.00 USD (or less) for 20-minute accounting tasks (e.g., Stinson, Doxey, and Rupert 2017; Morrow, Stinson, and Doxey 2018), which translates to a total cost of $2.80 based on a 40 percent markup, can attract hundreds of workers within one day. By contrast, Stinson, Barnes, Buchheit, and Morrow (2018) pay a flat rate of US$7.10 per participant to Qualtrics for a comparable task using a "Qualtrics Panel" of participants. Qualtrics does not disclose the portion actually remitted to participants. Other online participant sources claim to offer highly specialized samples, which generally come at a higher cost. For example, the RAND American Life Panel, which maintains a demographically representative panel of the U.S. working population, currently charges a fixed fee of $2,000 USD plus an additional cost of up to $3.00 *per minute* (based on RAND's estimated average completion time prior to launch) for each participant. Despite this heightened cost to the researcher, RAND's participants are typically paid a flat rate of $20 for a 30-minute survey.

[7] To illustrate a common request, MTurk workers occasionally experience "timeouts" if they take too long to complete a task and/or leave their web browser open for an extended period. Such disruption can prevent workers from completing a task or successfully requesting payment within the original HIT. Some workers request payment via email after explaining the situation. Based on the explanation and the record of work completed, researchers can choose whether to accommodate these requests.

[8] To date, TurkPrime has more than 1,700 registered users who have run more than 9,000 unique HITs in MTurk, garnering responses from over 60,000 unique workers (Litman et al. 2017). The TurkPrime website, https://www.turkprime.com, offers short video tutorials and a number of helpful tips.

[9] Litman et al. (2017, 434) specifically focused their efforts on "improving functionality over MTurk in six general areas: control over who participates in the study, flexible control over running HITs, more flexible communication and payment mechanisms, tools for longitudinal and panel studies, tools to increase sample representativeness, and enhanced study flow indicators."

[10] Blocks and rejections can be quite costly to workers—Amazon tracks these activities and does not require detailed feedback that might offer clarification of possible errors, instead reserving the right to terminate a worker's account at the company's discretion. Thus, while workers generally argue that blocks and rejections should only be used for egregious cases of bad faith and/or ineptitude, a researcher may be left with few attractive alternatives if he/she does not wish to pay for responses deemed unusable for more mundane reasons.

Given that accounting research often includes performance-based pay (e.g., Brink et al. 2017; Herschung, Mahlendorf, and Weber 2018), TurkPrime's payment features offer important benefits to many accounting researchers. To explain, the "bonus" feature in MTurk's graphical interface does not allow easy batch processing of additional compensation. Thus, a requester would typically have to grant variable or discretionary compensation (e.g., performance bonuses) by searching for each worker's unique ID number. In contrast, TurkPrime provides a more flexible bonus tool that allows researchers to grant equal and simultaneous bonuses to multiple participants in a study. For example, a requester could easily rank participants' relative performance and grant bonuses to segments of the distribution in bulk (e.g., a different bonus amount for each quartile of the sample).

While TurkPrime currently offers its "core" features (including those described above) to academic researchers at no additional cost, there are also several professional features available for an added fee.[11] For example, TurkPrime offers a "hyperbatch" function that automatically breaks the total sample size into multiple tasks of less than ten workers, thereby securing Amazon's minimum 20 percent markup rather than the 40 percent applicable to single surveys of ten or more.[12] This feature effectively pays for itself since the incremental fee to "hyperbatch" is less than the avoided markup to MTurk. In addition, TurkPrime's professional features allow requesters to target select geographic locations (e.g., by state or country), block duplicate IP addresses, and create survey groups that prevent workers from joining more than one HIT from a larger collection specified by the requester. Thus, while TurkPrime offers free enhancements to the native MTurk environment for academic researchers, it also provides a broad menu of premium services that can further expedite data collection at a low cost.

ATTRACTING MTURK PARTICIPANTS AND DISCLOSING SAMPLE SELECTION TECHNIQUES

In this section, we discuss various methods for screening participants and disclosing sample selection techniques that are particularly important for studies using MTurk workers or other crowdsourced participants.

Participant Screening Questions

Paramount to any behavioral study is sampling from the "right" subject population, and a successful sample collection relies heavily on the researcher's screening criteria and instrument design. For instance, it is common for researchers to sample distinct demographic groups based on subjects' age, employment status, marital status, etc.[13] However, since researchers must often communicate these requirements to participants before data collection under Institutional Review Board (IRB) policies, researchers cannot entirely rule out the possibility that attentive subjects may gain access to an instrument by providing false answers that comply with stated selection criteria.[14] This risk may be particularly acute for a "default" sequence in which a researcher asks screening questions immediately after any required IRB disclosures and before the experimental task.

Researchers can limit such opportunistic behavior in a number of ways, such as using advance criteria like MTurk's "premium qualifications" to avoid obvious screening questions. Alternatively, a researcher could establish a lengthy "break" between IRB disclosures and screens by allowing all participants to complete a substantial portion of the instrument before responding to screening questions. However, by increasing the number of responses collected before reaching selection criteria, the researcher faces increasing pressure to avoid late terminations or rejections and pay each participant the full HIT fee

[11] Currently, the cost to use each of TurkPrime's professional features is $0.02 per participant plus 5 percent of each participant's payment. These amounts are paid directly to TurkPrime in addition to the fees assessed by MTurk.

[12] The "hyperbatch" function is deemed a professional feature because it allows multiple HITs of smaller increments (e.g., ten HITs limited to nine workers each) to be launched simultaneously, thereby mimicking the immediate launch of a single, larger HIT (e.g., one HIT limited to 90 workers). Alternatively, TurkPrime offers a free "microbatch" feature that breaks larger worker quotas into smaller increments and spreads them over regular time intervals specified by the requester (e.g., ten HITs limited to nine workers each, launched at a rate of one HIT per hour for ten hours). While this method may lengthen the data-collection process, the "microbatch" feature can often secure greater cost savings than "hyperbatching" and perhaps alleviate concerns a researcher may have about potential bias or disparity among the participants available during peak or off-peak times of day (e.g., evening versus normal business hours).

[13] As a practical matter, we also recommend requesting a worker's unique MTurk ID in the early stages of an instrument to ensure payment applications can be matched to completed responses (and duplicate responses can be identified by something other than an IP address) and using at least one "Captcha" question. "Captcha" questions are a common feature of survey platforms (e.g., Qualtrics) that use randomly generated and generally unsearchable images to request input verifying that the user is a real person rather than a "robot" algorithm sometimes used in online environments. Examples include using images of street signs to request address information and/or having a user select multiple images on a screen that share a common characteristic (e.g., "from the following, please select every image of food").

[14] In the event participants do not read such IRB disclosures closely, the researcher must further decide if he/she wants to prohibit reentry to an instrument (such as by blocking duplicate IP addresses) after a failed screening question or reject participants who attempt an instrument more than once and vary their answers to screening questions before applying for payment. Unfortunately, both of these options are likely to inconvenience any participants who make honest mistakes in their initial screening responses and attempt to correct them.

regardless of qualifications.[15] Yet another possible approach keeps the screening questions near the beginning of the instrument but reduces demand effects by presenting subjects with a wide number of specific response options, only some (or one) of which meets the participation requirements.[16] This would seemingly reduce demand effects by making the "right" choice less transparent and less subject to guessing.

Finally, the speed and relatively low cost with which researchers can launch HITs affords them a unique opportunity to cultivate their own subject pools using a multistage approach, particularly when researchers target HITs with the invitation-only features available through third-party services such as TurkPrime. In the first stage, the researcher can launch a HIT featuring any number of potential screening questions, presumably without an accompanying experimental task or any explicit subject requirements contained in IRB disclosures. In other words, the first stage can be a general-purpose survey that offers little insight into the researcher's current agenda and provides a guaranteed payment with no obvious incentive for workers to tailor their responses. For a short survey consisting of demographic questions or other indicators of general ability (e.g., the Raven's Progressive Matrix questions used by Buchheit et al. [2017] to examine online participants' problem-solving abilities), the researcher could expect to pay a low fixed fee for all completed HITs. In the second stage, the researcher could then use an invitation-only HIT (e.g., through TurkPrime) to target participants whose answers in the first stage meet the screening criteria. As well as creating a longer "break" between screening questions and the primary task, this approach lowers the number of questions required in the second stage, thus reducing the time needed to complete the primary instrument and lowering associated risks of subject distraction or fatigue.

Screening Specific Populations of Interest in Accounting

While the previous section explains several ways in which researchers can alter the general structure of participant screening measures to enhance data integrity and improve data-collection efficiency, it is also possible to recruit a desirable population by leveraging workers' performance ratings and calibrating the content of screening questions. For example, Bentley (2017) lists several studies that screen participants based on their approval ratings in MTurk (Peer, Vosgerau, and Acquisti 2014), ability to pass instructional manipulation checks (Oppenheimer et al. 2009), and number of accepted HITs (Peer et al. 2014). Similarly, researchers can use screening questions to verify participants' expertise in a specific subject.[17] In this section, we offer examples of screening techniques used in recent works to attract populations of particular interest to accounting. Specifically, we focus our attention on studies that seek to recruit potential jurors, taxpayers, and nonprofessional investors.[18]

Multiple studies have found MTurk workers to be suitable proxies for jurors (e.g., Grenier, Pomeroy, and Stern 2015b; Grenier, Lowe, Reffett, and Warne 2015a; Brasel, Doxey, Grenier, and Reffett 2016; Maksymov and Nelson 2017).[19] For example, in addition to requiring participants to be U.S. citizens and at least 18 years old, Brasel et al. (2016) limited their sample to participants who had never served on a jury or worked in a field related to those involved in their experimental case. Maksymov and Nelson (2017) further list proficiency in English among their screening criteria, citing the qualifications for jury service on U.S. courts. Overall, these "juror studies" share a common theme in using a combination of general and case-specific criteria to eliminate respondents who would typically be eliminated in *voir dire* for real-world court proceedings (Brasel et al. 2016).

Similarly, experimental research on tax policy and compliance typically requires participants with at least modest tax-filing experience.[20] For example, Brink and Lee (2015) study the effects of refund/amount due status bars, commonly used in tax software packages, on decisions to report cash income among MTurk workers having filed six or more years of U.S. federal income tax returns. Stinson et al. (2017) and Cuccia, Doxey, and Stinson (2017) both examine the effects of Roth and traditional tax structures on retirement savings decisions. In both studies, the authors limit their samples to MTurk workers who report a history of tax filings and can identify both their most recent tax filing status (e.g., single versus married filing jointly) and their

[15] As discussed below, researchers face administrative and reputational costs from participants who feel they have been treated unfairly by screening mechanisms.

[16] For example, Doxey et al. (2017) disclose that they desire "non-professional investors" as participants. As a screening question, the authors asked participants which of ten different asset types they invest in, and only retained those participants who chose "stock in individual companies" as one of their responses.

[17] If researchers want a particular kind of expertise, then they can ask pointed questions that only experts would be able to answer. In this manner, the risk of falsely claimed expertise is mitigated.

[18] While we focus on these groups for illustrative purposes, other targeted populations of interest certainly exist. For example, Rennekamp, Rupar, and Seybert (2015) examine managerial decisions using MTurk workers who report prior work experience in accounting, finance, or management.

[19] Grenier et al. (2018) provide an excellent summary of design and recruiting challenges unique to "juror studies."

[20] Such experience with tax filings need not convey tax expertise. Rather, research on tax policy and compliance often focuses on decisions made by current taxpayers, an inherently diverse group with respect to financial literacy, socioeconomic status, professional experience, etc. Studies focused on broad perceptions among the population of potential taxpayers may simply require random dispersion of demographic characteristics (e.g., age, income, gender) and request information on prior interactions with tax authorities to use as controls rather than explicit screening criteria (e.g., Farrar et al. 2018b; Farrar, Hausserman, Pinto 2018a).

typical method of return preparation (e.g., self-prepared versus hired professional). These studies further screen on age to ensure their samples align demographically with the decision period of interest. Cuccia et al. (2017) limit their participants to 19 to 59 years of age to exclude participants already eligible to take qualifying retirement distributions from existing plans, while Stinson et al. (2017) further restrict their sample to ages between 28 and 43 to maintain similar distance from "end game" retirement investing patterns potentially spurred by age restrictions on qualifying withdrawals and mandatory retirement distributions.[21]

Next, Libby et al. (2002) suggest that individuals who have at least a basic understanding of accounting and investing may be suitable subjects for studies requiring nonprofessional investors, and Elliott et al. (2007) demonstrate that convenience samples from graduate accounting and finance courses can be used to this effect. Owens (2014) extends both studies by replicating experimental tasks from Elliott et al. (2007) on prescreened MTurk workers who had completed two or more courses in accounting or finance. Owens (2014) finds mixed results suggesting that MTurk workers "do not always acquire information in the same way as investors, and in some regards, MTurk workers' acquisition is inferior to that of MBA students." Owens's (2014) results indicate that this subsample of MTurk workers has similar levels of financial literacy and work and investing experience, but has less experience in financial statement analysis than both M.B.A. students and investors. Thus, additional screening on such analytical abilities might further bridge the gap between MTurk workers and more traditional samples. Similarly, Krische (2015) finds that the broad population of MTurk workers may require further screening on investment experience and numerical skills to obtain samples comparable to other M.B.A. and nonprofessional investor pools. Overall, these studies highlight that research using MTurk workers must strike an important balance between comparability to prior studies and generalizability to a wider population. While heightened financial and numerical screening from such a broad population promotes comparability to prior studies using M.B.A. students and nonprofessional investors, it likely sacrifices generalizability to a wider cross-section of "real world" market participants who vary considerably in their investing prowess and sophistication.[22]

Comprehension Checks

Like screening questions, the design of comprehension checks affords multiple options, each with its own potential strengths and weaknesses. Since many screening features can also be applied to comprehension checks, we focus primarily on differences between the two measures that may further impact research design. Of primary importance is the fact that, unlike screening questions, comprehension checks need not be viewed as a "one shot" game whereby initial failure precludes a subject's inclusion in a sample.[23]

For instance, comprehension checks can temper the ill effects of MTurk "speeders" (i.e., those who "race through" an online survey) (Smith et al. 2016) if researchers use them to immediately terminate a response session.[24] If researchers base comprehension checks on randomly assigned elements of the research design, then they may be less susceptible to repeated guessing, and "speeders" may choose to increase their attention to the task or abandon it altogether in favor of simpler HITs.

Although researchers should tie some comprehension checks to specific design elements, we also recommend having at least one check dedicated to payment parameters, which are particularly important to many MTurk workers. It is not uncommon for MTurk workers to contact researchers electronically to express displeasure after being denied payment. Thus, researchers can reduce potential conflicts with workers over rejected or screened responses by including comprehension checks detailing the amount of work needed to secure payment for a task (e.g., completing all questions on the task and successfully answering comprehension checks). Similarly, requiring participants to accurately identify both fixed and variable (or "bonus") compensation parameters ensures participants understand an experiment's incentives and prevents complaints over payment amounts at the end of the study.

---

[21] Morrow et al. (2018) further demonstrate that age may be a reasonable proxy for tax experience and understanding, as MTurk participants under 35 demonstrate an economically irrational bias toward tax incentives expressed as relatively low-transparency changes in personal deductions over mathematically equivalent tax credits and surcharges. The disparity in the authors' observed responses to tax incentives targeting personal health insurance between participant groups above and below age 35 align closely with recent IRS Statistics of Income data showing that younger taxpayers account for a disproportionately low share of income and complex tax positions (e.g., itemized deductions) relative to their 35 and older counterparts.

[22] Other studies have taken a relatively simpler approach to identifying nonprofessional investors in MTurk. As described above, Doxey et al. (2017) asked initial respondents to select their current investment vehicles from a list of several common vehicles (IRA, 401(k), etc.) and retained participants who indicated they had directly invested in a company's stock. Similarly, Rennekamp (2012) examines MTurk workers who report direct stock investment or an intention to invest directly in individual stocks in the future.

[23] On the other hand, some studies use comprehension checks as an additional screening tool to avoid paying for inattentive responses. For example, Brasel et al. (2016) rejected payment for participants who completed the study but did not correctly answer at least 90 percent of the comprehension checks included throughout the instrument. An evaluation of the accounting literature reveals no consensus regarding compensating participants who fail comprehension checks and therefore fail to complete the experiment. We observe researchers who provide full payment, partial payment, or no payment to such participants.

[24] Oppenheimer et al. (2009) also recommend the use of instructional manipulation checks, which embed specific instructions (e.g., click a hidden link at the top of the screen) within larger questions appearing to request another form of input (e.g., Which of these activities do you engage in regularly? Click all that apply.). The authors suggest these questions can increase the statistical power and reliability of a dataset by detecting participants who are not diligent in reading and following instructions.

Potential Administrative and Reputational Costs for Researchers Using MTurk

While much of the preceding discussion offers guidance for structuring MTurk HITs to minimize sources of friction or confusion between the researcher and workers, unforeseen issues may arise during data collection for any number of reasons. While these issues will vary in severity and the extent to which experimenters can control them, MTurk workers generally hold the researcher accountable for any perceived or experienced difficulties in completing a HIT.[25] To the extent a researcher is unable or unwilling to honor any special requests from workers, the researcher may face additional reputational costs in the form of negative participant reviews. Such reviews are generally communicated in online evaluations, such as those maintained by Turkopticon (which restricts its membership and often prevents researchers from responding directly to workers' concerns), and can discourage other workers from attempting a requester's tasks depending on the severity of reported claims. Workers may similarly use online evaluations to warn others of perceived inadequacies in a requester's HITs, such as a lack of clarity in the instrument, a task length that is incommensurate with the payment offered, or a requester's aggressive use of participant rejections and blocks.

Suggested Disclosures

One potential benefit of crowdsourcing is that theories can be more rapidly developed, tested, and fine-tuned, potentially speeding progress in psychology and the social sciences (Mason and Suri 2012; Chandler and Shapiro 2016). In contrast, some researchers in accounting (e.g., Libby et al. 2002; Trotman 2016) caution that this aspect of online data collection could work against scientific rigor, perhaps causing researchers to lose sight of careful design and theory development in favor of more stylized techniques that produce convenient results.[26] Appropriate and transparent disclosures might alleviate these concerns, but we have observed significant variation in disclosure, both in published and unpublished accounting papers.[27] Since we cannot observe details communicated to reviewers and editors outside of manuscripts, we caution against drawing inferences about the appropriateness of disclosures in prior and current literature. Rather, in this section, we offer some suggestions for disclosures that might foster more conformity in future accounting studies using MTurk and curtail any perceptions of strategic behavior.

First, we discuss general disclosures that may be informative to readers. Since MTurk workers have considerable freedom in choosing tasks from a wide assortment of HITs, it may be informative to disclose the posted HIT description. Other helpful disclosures may include requested worker qualifications and any methods used to set worker expectations, such as the estimated completion time advertised to workers. Similarly, observed completion times and any fixed or variable payment parameters necessary to calculate an effective hourly wage may be informative, particularly when task performance is an essential component of an experiment.[28] Finally, in certain research settings, the time the HIT remained active before reaching the desired number of workers could be useful, as it could provide a measure of the overall interest in the topic.

We next consider disclosures associated with sample attrition. Broadly speaking, it is reasonable to expect researchers to report any methods used to screen participants engaged in common forms of "bad" behavior (e.g., participants trying to speed or cheat their way through an instrument). In classic experiments with professionals or students, reporting such sample size reductions is often relatively straightforward and it is generally sufficient to report the number of manipulation failures or statistical outliers excluded from analysis. However, sample size and sample reduction considerations are different in an online recruitment environment. For example, unlike student populations taking part in an experiment (e.g., classroom recruiting), there is no accessible list of active potential MTurk workers at the time a HIT is released (Steelman et al. 2014). In addition, some participants agree to participate in an online survey, yet do little more than start the experiment before exiting. These issues beg the question of what steps in the sample selection process researchers should emphasize for the reader.

---

[25] For instance, the authors have encountered several participants who demand detailed explanations of their screening and/or comprehension check failures, but are seldom satisfied with the rationale. In the case of "timeouts," where a worker's access to MTurk is terminated during an extended browsing session that prevents him or her from applying for payment, the researcher generally has to establish a new HIT to pay affected workers. The authors find this process is well suited for the invitation features in TurkPrime, which allow a requester to limit a HIT's access to specified worker IDs and send email notifications when a new "payment HIT" (generally comprised of a single question) is available.

[26] For example, strategies such as recruiting larger sample sizes that afford more leeway in sample reductions and "tape spinning" by iterating instruments become easier and cheaper with online data collection.

[27] For example, approximately 33 percent of the papers published in premier accounting journals from 2010 through 2016 (see footnote 2) that used MTurk participants merely disclose the use of MTurk and/or provide basic demographic sample descriptions. Another 42 percent provide further details of sample attrition, such as failed attention or manipulation checks. The remaining 25 percent describe additional procedures performed to address challenges unique to using online participants (e.g., Koonce, Miller, and Winchel 2015; van der Heijden 2013; Zahller, Arnold, and Roberts 2015).

[28] Building on our discussion in the previous footnote, we also observe variation in the disclosures regarding participant compensation. For the MTurk papers published in premier accounting journals from 2010 to 2016, nearly two-thirds disclose the rate of pay for complete responses. The remaining third also include details regarding partial payments (e.g., in the case of early exits due to screening questions) or payment rejections (e.g., for failed attention checks).

To illustrate, researchers typically use software such as Qualtrics for expedience when collecting data through MTurk. When a potential participant accepts a HIT and clicks on the survey link, Qualtrics automatically increments a built-in observation count. However, the participant may decline to participate after reading initial IRB disclosures, fail screening questions (i.e., they were not members of the target population for sampling), quit the instrument before completion, or fail early comprehension checks (depending on a study's structure). The extent to which "early-exit" participants exist when using an online platform likely merits further investigation and perhaps disclosure.

When deciding whether to report the total number of participants "touching" (i.e., clicking a survey link) a protocol, researchers should consider whether there is something about their instrument that is causing systematic attrition. For example, a high percentage of "early-exit" participants could be indicative of an unclear or confusing instrument, or disinterest in the topic. In such cases, it may be necessary to report dropout rates by condition (Crump et al. 2013), with some researchers recommending that authors always disclose attrition rates for each experimental condition (e.g., Chandler and Shapiro 2016). Absent a significant number of "early-exit" participants, the more meaningful starting point for disclosing sample reductions is likely the number of participants who completed the instrument.

Considering complete responses as a starting point, a second sample reduction is often necessary to remove obviously invalid responses. These could include multiple responses from one IP address, implausibly short response times, respondents providing nonsense answers (e.g., numerical responses to qualitative questions), or any complete response from an IP address where the respondent clearly changed answers to screening questions after a prior rejection. Finally, reporting additional sample reductions for back-end attention or comprehension check failures (any not associated with screening questions), manipulation check failures, and outlier responses (if applicable) is also meaningful and desirable, just as with a sample of participants from a more traditional experimental setting.

To summarize, online platforms facilitate evaluation of sample attrition from the HIT acceptance until completion of the instrument. To ease any concern of systematic attrition that influences results, researchers could consider reporting "early-exit" attrition. Depending on the degree to which sample attrition could bias findings, researchers could disclose the following sample totals, either in direct communications to reviewers or within a manuscript: (1) everyone who clicked the survey link, (2) everyone who passed screening demographic requirements, (3) everyone who passed early attention and comprehension checks, (4) everyone who completed the instrument, (5) complete responses less "ballot stuffers" who circumvented screening criteria with repeated attempts, and finally (6) the remaining legitimate responses minus non-screening attention, manipulation, and outlier responses. Barring unusual circumstances, the most meaningful disclosures are likely the number of participants who completed the instrument and the corresponding sample attrition to the final sample size used in the reported statistical analyses (i.e., the cuts from subset four to six). When in doubt, we advise researchers to err on the side of over disclosure and be prepared to report and explain all attrition to reviewers and editors.

Finally, due to the nature of MTurk, the phrase "pilot testing" may be ambiguous. For example, it is common and prudent for researchers to run a "soft opening" by releasing an experiment to a small group of participants to ensure it works as expected and participants can complete the instrument. Soft openings can be particularly beneficial for experiments with complex survey coding and may also help refine completion time expectations. If the soft opening works and the researchers make no changes to the instrument, then it is appropriate to include the soft opening sample in the full analysis. This differs from iterative hypothesis testing where researchers run a full-scale experiment, analyze the results, and then make changes to the instrument. We propose that researchers briefly disclose the number and results of all pilot tests where they tested their hypotheses, as well as any changes to the instrument that result. Disclosing this information has the potential to limit concerns of "tape spinning" in MTurk.

CONCLUSION

The purpose of this article is to aid accounting researchers in using Amazon's Mechanical Turk (MTurk) online data-collection platform. MTurk use has grown exponentially in the behavioral sciences, raising concerns about how well MTurk research participants proxy for traditional research participants (Chandler et al. 2014). Of general use to behavioral researchers (and reviewers), our review of recent research (e.g., Chandler and Shapiro 2016) suggests that concerns about potentially low-quality MTurk participants are overstated. Specifically, MTurk participants are reasonable proxies for nonprofessional subjects with limited exceptions (e.g., non-U.S. MTurk participants may be cause for concern). For researchers considering experiments involving variable bonus payments or panel data, we discuss how cumbersome aspects of the native MTurk platform can be overcome with third-party services such as TurkPrime at a low (and potentially money-saving) cost. Finally, we discuss challenges associated with recruiting MTurk participants and how researchers might overcome such challenges and appropriately report efforts to attract participants and analyze responses.

With respect to disclosures, the unique details of the experiment likely dictate what information researchers provide to readers as opposed to just reviewers, although researchers should generally disclose basic information such as the HIT

description, payment structures, and completion times. When recruiting subjects with specialized qualifications, disclosure of careful and thoughtful screening techniques may also be necessary. Regardless of one's target population, consistent and transparent disclosures may also go a long way toward alleviating concerns regarding data quality and integrity in the peer-review process. Similarly, we suggest that disclosing screening methods, methods to reduce or eliminate common forms of bad-faith responses, and sample reductions should be common practice in future studies using MTurk.

Overall, we find Amazon's MTurk offers several distinct features that may prove useful for behavioral accounting research. MTurk provides a large and diverse population of low-cost research participants; however, researchers should be cognizant of the associated risks related to recruiting anonymous participants. Finally, we note that careful planning and transparent disclosures can mitigate the risks associated with online subject recruiting, making MTurk a highly effective research tool.

REFERENCES

Ashton, R., and S. Kramer. 1980. Students as surrogates in behavioral accounting research: Some evidence. *Journal of Accounting Research* 18 (1): 1–15. https://doi.org/10.2307/2490389

Bentley, J. 2017. *Challenges with Amazon Mechanical Turk Research in Accounting*. Working paper, University of Massachusetts Amherst.

Berinsky, A. J., G. A. Huber, and G. S. Lenz. 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

Brandon, D. M., J. H. Long, T. M. Loraas, J. Mueller-Phillips, and B. Vansant. 2014. Online instrument delivery and participant recruitment services: Emerging opportunities for behavioral accounting research. *Behavioral Research in Accounting* 26 (1): 1–23. https://doi.org/10.2308/bria-50651

Brasel, K., M. Doxey, J. Grenier, and A. Reffett. 2016. Risk disclosure preceding negative outcomes: The effects of reporting critical audit matters on judgments of auditor liability. *The Accounting Review* 91 (5): 1345–1362. https://doi.org/10.2308/accr-51380

Brink, W., and L. Lee. 2015. The effect of tax preparation software on tax compliance: A research note. *Behavioral Research in Accounting* 27 (1): 121–135. https://doi.org/10.2308/bria-50977

Brink, W., T. Eaton, J. Grenier, and A. Reffett. 2017. Deterring unethical behavior in online labor markets. *Journal of Business Ethics* (May): 1–18. doi:10.1007/s10551-017-3570-y

Buchheit, S., D. Dalton, T. Pollard, and S. Stinson. 2017. *How Smart Are Online Workers? A Student versus MTurk Participant Comparison*. Working paper, The University of Alabama and Clemson University.

Chandler, J., and D. Shapiro. 2016. Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology* 12 (1): 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Chandler, J., P. Mueller, and G. Paolacci. 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

Chilton, L., J. Horton, R. Miller, and S. Azenkot. 2010. *Task Search in a Human Computation Market*. Proceedings of the ACM SIGKDD Workshop on Human Computation, Washington, DC, July 25.

Crump, M., J. McDonnell, and T. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8 (3): e57410.

Cuccia, A., M. Doxey, and S. Stinson. 2017. *The Relative Effects of Economic and Non-Economic Factors on Taxpayers' Preferences between Front-Loaded and Back-Loaded Retirement Savings Plans*. Working paper, The University of Oklahoma and The University of Alabama.

Doxey, M., R. Hatfield, J. Rippy, and K. Peel. 2017. *Asymmetric Investor Materiality: The Effects of Gains, Losses, and Disclosures*. Working paper, The University of Alabama and Florida State University.

Elliott, W. B., F. D. Hodge, J. J. Kennedy, and M. Pronk. 2007. Are M.B.A. students a good proxy for nonprofessional investors? *The Accounting Review* 82 (1): 139–168. https://doi.org/10.2308/accr.2007.82.1.139

Farrar, J., C. Hausserman, and O. Pinto. 2018a. *Trust and Compliance Effects of Taxpayer Identity Theft: A Moderated Mediation Analysis*. Working paper, Ryerson University.

Farrar, J., S. Kaplan, and L. Thorne. 2018b. The effect of interactional fairness and detection on taxpayers' compliance intentions. *Journal of Business Ethics* (forthcoming). https://doi.org/10.1007/s10551-017-3458-x

Farrell, A., J. Grenier, and J. Leiby. 2017. Scoundrels or stars? Theory and evidence on the quality of workers in online labor markets. *The Accounting Review* 92 (1): 93–114. https://doi.org/10.2308/accr-51447

Grenier, J., D. Lowe, A. Reffett, and R. Warne. 2015a. The effects of independent expert recommendations on juror judgments of auditor negligence. *Auditing: A Journal of Practice & Theory* 34 (4): 157–170. https://doi.org/10.2308/ajpt-51064

Grenier, J., B. Pomeroy, and M. Stern. 2015b. The effects of accounting standard precision, auditor task expertise, and judgment frameworks on audit firm litigation exposure. *Contemporary Accounting Research* 32 (1): 336–357. https://doi.org/10.1111/1911-3846.12092

Grenier, J., A. Reffett, C. Simon, and R. Warne. 2018. Researching juror judgment and decision making in cases of alleged auditor negligence: A toolkit for new scholars. *Behavioral Research in Accounting* 30 (1). https://doi.org/10.2308/bria-51878

Gureckis, T., J. Martin, J. McDonnell, A. Rich, D. Markant, A. Coenen, D. Halpern, J. Hamrick, and P. Chan. 2016. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods* 48 (3): 829–842. https://doi.org/10.3758/s13428-015-0642-8

Harms, P. D., and J. A. DeSimone. 2015. Caution! MTurk workers ahead—Fines doubled. *Industrial and Organizational Psychology: Perspectives on Science and Practice* 8 (2): 183–190. https://doi.org/10.1017/iop.2015.23

Hauser, D., and N. Schwarz. 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

Herschung, F., M. Mahlendorf, and J. Weber. 2018. Mapping quantitative management accounting research 2002–2012. *Journal of Management Accounting Research* 30 (1). https://doi.org/10.2308/jmar-51745

Horton, J., D. Rand, and R. Zeckhauser. 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. 2010. *Demographics of Mechanical Turk*. Working paper, New York University.

Koonce, L., J. Miller, and J. Winchel. 2015. The effects of norms on investor reactions to derivative use. *Contemporary Accounting Research* 32 (4): 1529–1554. https://doi.org/10.1111/1911-3846.12118

Kraut, R., J. Olson, M. Banaji, A. Bruckman, J. Cohen, and M. Couper. 2004. Psychological research online. *The American Psychologist* 59 (2): 105–117. https://doi.org/10.1037/0003-066X.59.2.105

Krische, S. 2015. *The Impact of Individual Investors' Financial Literacy on Assessments of Conflicts of Interest*. Working paper, American University.

Lease, M., J. Hullman, J. Bigham, M. Bernstein, J. Kim, W. Lasecki, S. Bakhshi, T. Mitra, and R. Miller. 2013. *Mechanical Turk Is Not Anonymous*. Working paper, The University of Texas.

Libby, R., R. Bloomfield, and M. Nelson. 2002. Experimental research in financial accounting. *Accounting, Organizations and Society* 27 (8): 775–810. https://doi.org/10.1016/S0361-3682(01)00011-3

Litman, L., J. Robinson, and T. Abberbock. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49 (2): 433–442. https://doi.org/10.3758/s13428-016-0727-z

Maksymov, E., and M. Nelson. 2017. Malleable standards of care required by jurors when assessing auditor negligence. *The Accounting Review* 92 (1): 165–181. https://doi.org/10.2308/accr-51427

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

Morrow, M., S. Stinson, and M. Doxey. 2018. Tax incentives and target demographics: Are tax incentives effective in the health insurance market? *Behavioral Research in Accounting* 30 (1).

Oppenheimer, D. M., T. Meyvis, and N. Davidenko. 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

Owens, J. 2014. *Using Mechanical Turk (MTurk) Workers for Nonprofessional Investor Research*. Working paper, University of South Carolina.

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

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

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

Rennekamp, K. 2012. Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research* 50 (5): 1319–1354. https://doi.org/10.1111/j.1475-679X.2012.00460.x

Rennekamp, K., K. Rupar, and N. Seybert. 2015. Impaired judgment: The effects of asset impairment reversibility and cognitive dissonance on future investment. *The Accounting Review* 90 (2): 739–759. https://doi.org/10.2308/accr-50879

Ross, J., L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. 2010. *Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk*. Proceedings of the ACM Conference on Human Factors in Computing Systems, New York, NY.

Smith, S., C. Roster, L. Golden, and G. Albaum. 2016. A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research* 69 (8): 3139–3148. https://doi.org/10.1016/j.jbusres.2015.12.002

Steelman, Z., B. Hammer, and M. Limayem. 2014. Data collection in the digital age: Innovative alternatives to student samples. *MIS Quarterly* 38 (2): 355–378. https://doi.org/10.25300/MISQ/2014/38.2.02

Stewart, N., C. Ungemach, A. Harris, D. Bartels, B. Newell, G. Paolacci, and J. Chandler. 2015. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making* 10 (5): 479–491.

Stinson, S., M. Doxey, and T. Rupert. 2017. *The Effects of Income Tax Timing and Performance Feedback on Retirement Investment Decisions*. Working paper, The University of Alabama.

Stinson, S., B. Barnes, S. Buchheit, and M. Morrow. 2018. Do consumer-directed tax credits effectively increase demand? Experimental evidence of conditional success. *The Journal of the American Taxation Association* (forthcoming).

Trotman, K. T. 2016. *Potential Pitfalls of Online Platforms*. Panel Presentation at the 2016 Research Conference of the Accounting, Behavior and Organizations Section of the American Accounting Association, Albuquerque, NM.

van der Heijden, H. 2013. Charities in competition: Effects of accounting information on donating adjustments. *Behavioral Research in Accounting* 25 (1): 1–13. https://doi.org/10.2308/bria-50295

Zahller, K., V. Arnold, and R. Roberts. 2015. Using CSR disclosure quality to develop social resilience to exogenous shocks: A test of investor perceptions. *Behavioral Research in Accounting* 27 (2): 155–177. https://doi.org/10.2308/bria-51118