

MTurk Character Misrepresentation: Assessment and Solutions

KATHRYN SHARPE WESSLING
JOEL HUBER
ODED NETZER

This tutorial provides evidence that character misrepresentation in survey screeners by Amazon Mechanical Turk Workers (“Turkers”) can substantially and significantly distort research findings. Using five studies, we demonstrate that a large proportion of respondents in paid MTurk studies claim a false identity, ownership, or activity in order to qualify for a study. The extent of misrepresentation can be unacceptably high, and the responses to subsequent questions can have little correspondence to responses from appropriately identified participants. We recommend a number of remedies to deal with the problem, largely involving strategies to take away the economic motive to misrepresent and to make it difficult for Turkers to recognize that a particular response will gain them access to a study. The major short-run solution involves a two-survey process that first asks respondents to identify their characteristics when there is no motive to deceive, and then limits the second survey to those who have passed this screen. The long-run recommendation involves building an ongoing MTurk participant pool (“panel”) that (1) continuously collects information that could be used to classify respondents, and (2) eliminates from the panel those who misrepresent themselves.

Keywords: Amazon Mechanical Turk, deception, panel, screener questions, theory-driven sample

Character misrepresentation occurs when a respondent deceitfully claims an identity, ownership, or behavior in order to qualify and be paid for completing a survey or a behavioral research study. For a large number of marketing studies, accurate screening is critical for the effective understanding of market behavior. [Goodman and Paolacci](#)

Kathryn Sharpe Wessling (wessling@wharton.upenn.edu) is visiting faculty at the Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104; Joel Huber (joel.huber@duke.edu) is a professor at the Fuqua School of Business at Duke University, 100 Fuqua Drive, Durham, NC 27708; and Oded Netzer (onetzer@gsb.columbia.edu) is an associate professor at the Graduate School of Business, Columbia University, 3022 Broadway, Uris Hall 520, New York, NY 10027. The authors gratefully acknowledge Gabriele Paolacci, Joseph Goodman, and JCR editors for their feedback and contributions to this tutorial. Supplementary materials are included in the web appendixes accompanying the online version of this tutorial.

This tutorial was invited by editors Darren Dahl, Eileen Fischer, Gita Johar, and Vicki Morwitz.

Advance Access publication April 17, 2017

(2017) articulate the need for theory-driven samples. For example, a study about uterine cancer treatment options makes little sense if it includes males.

Our own interest in this topic came from three experiences while engaging in research with Amazon Mechanical Turk (MTurk) participants:

- The authors needed a large number of respondents who frequented burger-related fast food restaurants at least once a month. Out of the 1,754 Turkers who passed a three-question screener, 149 did so by making multiple attempts at passing the screener questions. Another 100 made multiple attempts but were not able to figure out the combination of answers that would permit passage ([Wessling, Netzer, and Huber 2016](#)).¹
- The second author ran two conjoint studies seeking ways to help patients explore and communicate their wants and needs with their physicians. Smokers over 50 qualified for a study of lung cancer treatments, while active athletes under 35 qualified for a study of shoulder dislocation

1 While the “Prevent Ballot Box Stuffing” option was selected in this Qualtrics study, participants can make multiple attempts at a study if they clear the cookies from their web browser or simply switch browsers.

treatments. Seventeen percent of respondents in the cancer study had the same Worker IDs as those in the shoulder study (Tong et al. 2012).²

- The third author asked for Turkers who had written over 10 reviews on Yelp to complete a study. Almost 900 Turkers began the study and all but 33 dropped out when they were asked to provide a screenshot that verified their qualifications.

These disturbing examples mirror similar cases reported by Chandler and Paolacci (2017) demonstrating consistent distortions in responses when MTurk participants are able to retake a screener or falsify their identities in order to complete a study. Our goal is to identify the degree of misrepresentation in paid MTurk studies and its implications on the legitimacy of the scientific inquiry. We then propose a two-step process to achieve appropriate theory-driven samples. The first step assesses a respondent's qualification in a context where the respondent has neither the motive nor the requisite knowledge to deceive. The second step then makes the study available and viewable only to those who have qualified in the first step. Finally, we detail ways that this two-step method can be incorporated into a larger panel creation and management process that enables research with known and trusted MTurk respondents.

Amazon Mechanical Turk is the focus in this tutorial on misrepresentation because Turkers provide the dominant source of web-based studies for those studying consumer behavior (Goodman and Paolacci 2017). However, similar deception may occur on other crowdsourcing platforms, professional marketing research panels, or in-person studies. For example, a person interested in being a part of a focus group about diaper brands that pays \$150 may claim to be a mother with young children when in fact she is not (Leitch 2004). Thus, our recommendations are also relevant to other online and offline respondent recruiting platforms. While the problem is not limited to online studies, it may be particularly severe in this context given that one can more easily misrepresent oneself in the anonymity of an online environment.

There are four key lessons from this tutorial. First, we demonstrate that MTurk workers are willing to misrepresent themselves to gain access to a desired study, and that those who do so generate distorted responses to other questions in the study. Second, we show that the level of character misrepresentation is negligible when there is no economic motive to lie. Third, we characterize the role of online Turker communities, demonstrating how the goals of MTurk workers interact and sometimes conflict with the practices and values of the consumer behavior research community. Finally, we evaluate various measures to prevent misrepresentation, arguing that traditional measures of response quality are not very useful, but need to be replaced by a two-step process that separates the character identification from the study itself. Details on the mechanics are provided in the web appendixes.

There are a number of issues related to using MTurk respondents that are only briefly mentioned in this tutorial because they are well covered elsewhere. The important issue of the representativeness of the Turkers community to different populations has been extensively explored by other researchers (Berinsky, Huber, and Lenz 2012; Goodman and Paolacci 2017; Paolacci, Chandler, and Ipeirotis 2010; Ross et al. 2010). We also do not cover attrition rates due to study manipulations that can distort research conclusions, such as a writing task in one condition but not the other (Zhou and Fishbach 2016). Finally, we do not explore the disturbing finding that people who complete many social psychology research studies become non-naïve, and are thus differentially affected by specific manipulations, various forms of memory tasks, and attention checks (Chandler, Mueller, and Paolacci 2014; Chandler et al. 2015).

TESTING CHARACTER MISREPRESENTATION

We begin with a series of two-stage tests that assess the extent to which Turkers misrepresent themselves when they have a motive and opportunity to do so. In the first stage, respondents provide their demographic characteristics, activities, and product ownership in a context that does not offer any monetary incentive to misrepresent nor provides any information on the desired response. In the second stage, a screener question permits respondents to alter their answers from the first-stage questions in order to take a new study. Comparing respondents' answers across stages allows us to assess the degree of misrepresentation and the extent to which Turkers provide distorted answers to subsequent questions. We also compare these results to a simple take/retake group to separate misrepresentation from reliability in survey response.

Stage 1: Collecting Panel Characteristics

To assess character misrepresentation, we first built a panel with "true" characteristics and activities including product and pet ownership from 1,108 Turkers located in the United States. These questions were spread across eight different surveys that asked about (1) political and religious affiliations (MoralFoundations.org); (2) moral beliefs (MFQ: Graham et al. 2011); (3) material values (MVS: Richins 2004); (4) personality trait importance (GSA: adapted from Barriga et al. 2001); (5) extroversion and agreeableness (John and Srivastava 1999); (6) personality (TIPI: Gosling, Rentfrow, and Swann 2003); (7) product ownership (i.e., sports and technology), pet ownership (dog, fish, cat, etc.), food consumption (Sharpe, Staelin, and Huber 2008), health consciousness (Gould 1988), and social desirability bias (Crowne and Marlow 1960); and, (8) willingness to compromise moral beliefs (MFSS:

² A Worker ID is a unique identifier for each MTurk worker.

Haidt, Graham, and Joseph 2009). The specific contents of each survey are outlined in web appendix A; however, a thorough analysis of this data goes beyond the scope of this tutorial.

All eight surveys were launched simultaneously so that any MTurk worker could take as many surveys as desired within the first hour of posting. At the end of the hour, any worker who had taken one or more of the panel surveys became a “panelist” and gained access for the next four weeks to take any of the uncompleted eight surveys.³ Only those identified as panelists could see or take the panel surveys after the initial one-hour cutoff. On average, our panelists completed 7.1 panel surveys out of the eight available.

Each panelist saw a consent form at the beginning of each first- and second-stage survey. The consent form notified respondents of the possibility that their answers from other studies could be linked through their unique MTurk Worker ID, and if participants did not agree to these terms, they could exit the study. Including this consent form has implications, as respondents who expected to cheat may question whether they wanted to complete the survey or study, and thus, might drop out of our panel. However, we found the dropout rate to be minimal. Across eight surveys with more than 1,000 respondents, 96 respondents abandoned a survey, with only 16 of these occurring at the consent form stage.

Stage 2: Testing Misrepresentation

We conducted five studies to determine the extent to which participants altered earlier responses to qualify for a study. As detailed in web appendix B, the studies differed in terms of screening requirements and the questions asked in the body of the study. Only panelists were permitted to view the MTurk HIT (i.e., Human Intelligence Tasks) description and participate in the studies. In this second stage, the invitation described the general topic of the study (e.g., product-related study, health-related study, pet food survey) and whether it would be restricted to those with certain characteristics. We provided this detail to respondents for two reasons. First, in treating potential respondents ethically, analogous to many lab situations, we informed potential participants of the requirements so they could freely choose to take the study (Gleibs 2016). Second, because Turkers often complain about “unpaid” screeners, for four out of our five studies, we informed them of the qualification requirements a priori so they would not waste their time if they did not meet the requirement. If participants chose to accept the task, they clicked on a survey link and viewed the consent screen indicating that their responses

could be tied to other studies. Once respondents passed the screener and the study questions, they entered a unique completion code in order to be paid. Thus, our two-stage design allows us to assess the extent of misrepresentation when Turkers are given the opportunity to do so.

In discussing these studies, we concentrate on the degree of character misrepresentation and the distortion in responses to subsequent questions. We focus on responses that were statistically different between those who did and did not misrepresent themselves. Later sections examine the contexts in which strong misrepresentation occurs, the role of Turker communities and norms, and possible solutions to character misrepresentation in MTurk studies.

The five studies screened respondents on (1) owning a cat and/or a dog, (2) owning a kayak, (3) being over 49 years old, (4) being raised Catholic, and (5) being female. In all studies, we define impostors as those who provided the requested response to the screener question that differed from their response in stage 1. It is important to control for possible alternative explanations for inconsistent responses between the two stages, such as take/retake reliability error and change in status or character between the two surveys (e.g., someone may have purchased a kayak in between the two phases in our sports equipment–related study). We do so by including in four out of the five studies a “control” condition in which the “screener” question was included as part of the survey but not as a screener. The proportion of inconsistent responses between stage 1 and stage 2 in the control condition, where the focal question was not a screener, provides an estimate for differences that are due to random inconsistency or change in character status but not due to misrepresentation.

Table 1 provides for each study the percent of the first-stage panelists who had the qualification requirement when there was no incentive to lie in stage 1 (column A), and the percent of respondents in the second stage who altered their earlier response to enable them to take the study (column B). That shows unacceptable rates of misrepresentation ranging from 24% to 83%, with greater rates occurring when there are relatively few Turkers who can honestly respond to the screen (low rates in column A). Because the proportion of possible misrepresentation is “capped” at the proportion of respondents who are “eligible” to do so, we report (column C) the proportion of impostors (column B) divided by the proportion of respondents who are “eligible” to do so ($1 - \text{column A}$). This measure gives us a “standardized” degree of misrepresentation. Looking at column C, we see misrepresentation of around 80% for the pet and kayak ownership, but around 50% for age, religious upbringing, and gender. This suggests that respondents are less likely to deceive with respect to stable, identifiable demographic characteristics compared to product ownership, which is more difficult to disprove. We encourage future studies to further explore the kinds of

³ This was accomplished through using the MTurk qualification functionality. We created a qualification type called “qual” and set this value to 1 for every panelist (see the appendix for details). We also batch-notified our panelists of other surveys that they were eligible to take using the R package MTurkR, which may have contributed to the high response rate (see web appendix E).

TABLE 1
CHARACTER MISREPRESENTATION IN STUDIES WITH AND WITHOUT SCREENERS

Study	Qualification requirement in B	A: Panel survey: % of initial panel who satisfy screen (a)	B: Screened study: % of paid respondents who alter initial response to satisfy screen (b)	C: Deceivers: % of altered responses relative to those "eligible" to deceive (b)/(1 - (a))	D: Control: % of paid responses who alter response when there is no screen
Pet food study I	Must own a dog OR a cat	70% (n = 1,000)	24% (n = 378)	80%	NA
Pet food study II	Must own a dog AND a cat	19% (n = 1,000)	71% (n = 123)	88%	NA
Kayak study	Must own a kayak	7% (n = 1,000)	83% (n = 146)	89%	4% (n = 96)
Fiber study	Must be 50 years old or older	13% (n = 999)	43% (n = 141)	49%	0% (n = 144)
Politics study	Must have been raised Catholic	30% (n = 1,034)	39% (n = 120)	56%	4% (n = 138)
Cellphone case study	Must be female	49% (n = 1,041)	25% (n = 141)	49%	0% (n = 154)

screens that are more or less likely to encourage misrepresentation.

Column D gives the inconsistency rates in a control study where there was no screen and thus no motive to impersonate. We see a baseline inconsistency of 0–4% when there is no motive to deceive. That baseline inconsistency is important in providing the prime justification for screening in a separate survey.

We now describe each of the studies and the differences in responses between those who did and did not misrepresent themselves. Web appendix B provides the details of each of these studies.

Pet Ownership. We ran two tests related to pet food brands, with the first test requiring participants to have at least one dog *or* cat to qualify and the second test requiring at least one dog *and* one cat. Upon entering the second-stage tests, participants were asked to complete a screening question about pet ownership. If they reported having the required number (independent of whether they reported the correct answer in the first-stage survey), they were shown the consent screen and were permitted to take the study. Otherwise, they were told that they did not qualify and could not continue.

Examining table 1, 70% of the respondents indicated that they had either a dog or a cat in the first stage. In the second stage, 24% out of the 378 respondents who completed the study altered their earlier responses to gain access to the study. By contrast, for the more restrictive qualification, 19% of the responses in the initial survey indicated they had both kinds of pets, but in the second stage, 71% out of 123 respondents who completed the study changed their pet ownership response to qualify. Both levels of misrepresentation are unacceptable, but clearly the greatest risk occurs for the more restrictive screens.

Many of the subsidiary questions did not differ significantly between the respondents who misrepresented and those who did not. However, when given a list of 15 national brands of pet food and asked which one(s) they actually purchase for their pets, impostors were significantly more likely to claim that they purchase a national brand compared to the outside alternatives, either the “none” option or the store-branded food (dog food: 90% vs. 82%; $p = .033$; cat food: 94% vs. 84%; $p = .004$). Across our studies, we often found that impostors are significantly less likely to choose the “none” option. One possible explanation is that impostors want to appear knowledgeable and involved and hence are less likely to go beyond listed brands.

These results are disappointing in demonstrating substantial levels of misrepresentation and significant differences when it comes to study responses. While unlikely to explain the entire result, two possible explanations for the difference between the stage 1 and 2 pet ownership questions are changes in the pet ownership in the two months that passed between the two stages of the study, or take/re-take errors in response to the survey questions. To assess

the degree to which inconsistencies between the two studies may be attributed to such accounts, in the next studies we include a control group that received the same survey without any screeners. Responses from the control group also measure the fundamental variability in the response to the screening variable across stages.

Kayak Ownership. We determined kayak ownership in stage 1 by asking respondents about their sports equipment, ownership. In doing so, 7% of our panelists checked a box indicating that they currently owned at least one kayak. Thus, in this study, due to the relatively low ownership of kayaks reported in the first stage, 93% of the respondents to the first study had an opportunity to deceive. Two months later, a second-stage study was posted, stating it was just for kayak owners. Once past the consent screen, panel members chose again among the same sports equipment options as in stage 1 and were permitted into the paid study if they checked the box indicating that they owned a kayak. Of the 146 respondents in stage 2 who indicated that they currently owned a kayak, 132 (88%) had indicated earlier that they did not. However, seven participants also indicated that they had recently purchased a kayak, which leads us to conclude that at least 83% of stage 2 participants were clear kayak owner impostors.⁴

Because only 18 respondents reported both in the first and second stages that they owned a kayak, this study did not provide a sufficient sample size to compare the response of impostors and consistent respondents to other questions. In this study we asked a separate group of respondents to report their kayak ownership with no incentive to imposter (take/retake) and found that only 4% of those who reported to have a kayak in stage 2 did not report the same in stage 1. This may be due to the purchase of a kayak between the two studies (although no one indicated a recently purchased kayak) or due to response inconsistency. Thus, we can conclude that the vast majority of the change in response to the kayak ownership question between the two surveys is due to intentional misrepresentation and not merely inconsistency in response.

Dietary Fiber for Those over 50. In the first stage, 13% of panel respondents indicated that they were 50 years old or older. In the second stage, the recruiting statement explicitly stated that only those 50 and over would qualify. Upon entrance to the survey, participants viewed the consent screen and reported their age. Those who said they were 50 or above were permitted to take the study. There was substantial age misrepresentation, with 43% of the 141 stage 2 respondents being revealed as impostors. To make sure that the stage 2 age screen was not due to take/retake error, a separate group of panelists responded to a similar

survey but without any screener. 100% of the 144 respondents in this control condition reported an age bracket that was perfectly consistent with the age reported in stage 1.

Among other questions, participants made a choice of a fiber supplement among Metamucil Tablets (\$15.99), Fiber Well Gummies (\$14.99), Benefiber Powder (\$25.99), and a “none” option. The impostors, with an average age of 33, were significantly less likely to choose the “none” option relative to those who legitimately passed the screener (8% vs. 25%; $z = -2.567$, $p = .010$). Impostors also overstated their average vitamin intake frequency (ranging from never = 0 to daily = 3) compared to those legitimately over 49 years old ($M_{\text{impostors}} = 2.36$; $M_{>49} = 1.96$; $F(1,140)$, $p = .036$). Thus, we find that not only do respondents misrepresent their age, but more importantly, impostors exhibited different responses to other questions, leading to biased survey results.

Catholic Upbringing. In the first stage, 30% of panel members indicated that they had a Catholic upbringing. The second-stage recruiting statement specified that only those raised Catholic could take the study. Once in the survey, if respondents indicated in the screener question that they were not raised Catholic, the study ended and they were not compensated. However, if they claimed that they were raised Catholic, they completed the study and were paid regardless of whether their claim matched their first-stage response. Then participants were shown an excerpt from a CNN article (Burke 2016) reporting a controversy between Pope Francis and Donald Trump and asked if they agreed with the Pope’s statement that “A person who thinks only about building walls, wherever they may be, and not building bridges, is not Christian” (Strongly disagree = 1, Strongly agree = 5).

Of the stage 2 respondents, 61% of the 120 participants consistently matched their earlier statement that they had been raised Catholic, while the other 39% contradicted their earlier response about their religious upbringing. For comparison purposes, we relaunched the study with no screener, and only 4% of 138 respondents changed their reported religious upbringing in a take/retake study when there was no monetary incentive to misrepresent. Furthermore, we found that those raised Catholic were statistically more likely to agree with the Pope’s statement than the impostors ($M_{\text{Catholic}} = 3.93$; $M_{\text{impostors}} = 3.38$; $p = .028$).

Woman’s Cell Phone Case Conjoint. The final experiment tested gender misrepresentation and included a standard choice-based conjoint task. In the first four studies, the unscreened “control” condition was launched after the screening condition; thus, differences between control and screen may have been due to selection effects given that those who had previously taken the screener version of the study were excluded from taking the control relaunch. To mitigate such possible selection effects, we randomly assigned panel members either to a screen or no-screen condition, both of which were ran simultaneously. As shown in table 1, 25% of the 141 respondents in the screener

⁴ Note that respondents also had an incentive to lie about acquiring a kayak in between the studies to justify their inconsistency between the two studies.

FIGURE 1

EXAMPLE CHOICE TASK FROM THE CONJOINT EXERCISE FOR FEMALES

 <p>black</p> <p>OTTERBOX</p>	 <p>pink</p> <p>OTTERBOX</p>	 <p>navy</p> <p>Otter BOX</p>	<p>I would not choose any of these.</p>
<p>limited protection</p>  <p>EASY ON/OFF</p> <p>\$59.99</p>	 <p>DROP PROTECTION</p>  <p>RADIATION PROTECTION</p>  <p>SLIM DESIGN</p> <p>\$49.99</p>	 <p>DROP PROTECTION</p>  <p>ULTRA-SLIM PROILE</p> <p>\$29.99</p>	

condition changed their reported gender to gain entrance to the study. By contrast, none of the 154 respondents in the unscreened condition changed their gender identities.

All respondents completed 12 choice-based conjoint tasks selecting among cell phone case designs. As shown in an example task in figure 1, the attributes and levels for the alternatives included color (pink, black, or navy), style (slim design, ultra-slim profile, or easy on/off), drop protection (included or limited), radiation protection (included or limited), and price (ranging from \$29.99 to \$59.99).

Table 2 summarizes the conjoint estimates. We found that males posing as females statistically differed from true females on the stereotypically female attributes of color and design. Specifically, males impersonating as females had higher estimated utility (part-worth) for a pink cell phone case ($M_{\text{females}} = -0.53$; $M_{\text{impostors}} = 1.85$; $p = .013$) and an ultra-slim case profile ($M_{\text{females}} = 0.40$; $M_{\text{impostors}} = 1.09$;

$p < .0001$) compared to actual females surveyed. Those misrepresenting their gender also had a higher utility value for the “none” option ($M_{\text{females}} = -3.43$; $M_{\text{impostors}} = -1.70$; $p = .043$) and chose the “none” option more often than females ($M_{\text{females}} = 7\%$; $M_{\text{impostors}} = 13\%$; $p = .013$). This result may seem to contradict the earlier finding that impostors are less likely to choose the “none” option. However, when we examine the control condition, we can see that males posing as females had marginally lower utility values for the “none” option compared to males in the control condition. That result is consistent with our previous findings that those who impersonate tend to be more averse to choosing the “none” option compared to those who are being honest ($M_{\text{males}} = -0.06$; $M_{\text{impostors}} = -1.70$; $p = .088$). There was no reliable difference in utilities on the less stereotypically female attributes (i.e., drop and radiation protection) between males in the control condition and males posing as females.

TABLE 2

PARTWORTH UTILITIES FOR CELL PHONE CASE STUDY FOR IMPOSTOR MALES, FEMALES, AND NONIMPOSTOR/CONTROL MALES

Attribute level	Male impostors <i>n</i> = 35	Females <i>n</i> = 180	Males in the control condition <i>n</i> = 80
Color: pink (relative to a black) case	1.85 (SE = 1.00)	-0.53** (SE = 0.37)	-8.72*** (SE = 0.52)
Design: ultra-slim (relative to the easy on/off) case	1.09 (SE = 0.10)	0.40*** (SE = 0.06)	0.44*** (SE = 0.09)
Radiation protection: relative to not included	0.66 (SE = 0.10)	0.94** (SE = 0.05)	0.72 (SE = 0.06)
Drop protection: included relative to not included	2.18 (SE = 0.26)	2.85** (SE = 0.11)	2.29 (SE = 0.17)
The “none” option	-1.70 (SE = 0.76)	-3.43** (SE = 0.34)	-0.06* (SE = 0.53)

NOTE.—Difference with male impostors is significant
 ****p* < .001,
 **<.05,
 *<.1.

CONCLUSIONS FROM THE FIVE STUDIES

The five tests demonstrate that studies using screeners that rely on respondents’ self-reports are susceptible to an unacceptably large proportion of impostors. In particular, we find that from 24% to 83% of those passing the screener questions are impostors, and that deceit occurs in 49–89% of those who are “eligible” to misrepresent. The risk of misrepresentation is greater for narrow or rare screening categories and when the characteristic misrepresented is flexible, like ownership, rather than inflexible, like demographics. Thus, we can conclude that without safeguards, misrepresentation can be destructively common.

Further, those who pretend to be someone else may use one of three different strategies in answering questions. First, impersonators may be reluctant to admit their lack of knowledge and thus may be less likely to choose the “none” response. Second, impostors may attempt to project what they expect the mimicked persona would think, and in doing so overemphasize stereotypes. That appears to happen with male impostors improperly projecting that women prefer pink cell phone cases. Finally, where projection to a different person is difficult, deceivers may simply default to their own personal views or preferences. That may have happened when those misrepresenting their Catholic upbringing were more likely to disagree with the Pope than actual Catholics. The important point here is that there are various ways a deceiver may continue to deceive, and it is very difficult to predict the direction or magnitude of the bias.

The good news from our tests is the strong evidence of minimal distortion when there is no economic motive to do so. That occurred in the control studies having less than 5% inconsistency between the stages when there was no screener needed to gain entry into the study. This high degree of take/retake reliability among Turkers is reasonable,

simply because telling the truth is easy, while deceit takes effort. It also speaks to the fairly high internal validity of MTurk responses. Before we examine how one mitigates this threat to the validity of studies, it is important to understand the roles that web forums have on Turkers’ behavior and particularly on the likelihood of addressing deception.

ONLINE TURKER FORUMS AND DECEPTION

Given the substantial number of impostors in our test studies, we were interested in the potential role that online Turker communities have in either encouraging or discouraging deception. The following table provides a list of the major Turker forums.

A number of researchers have documented the frustration and difficulty associated with being a Turker (Dholakia 2015; Martin et al. 2014). MTurk online forums have been created by Turkers and serve four primary functions to limit that frustration. First, the websites help Turkers select desirable HITs by including estimates of actual pay per minute (which can differ from the estimated pay rate) and any warnings about difficult, boring (e.g., “bubble hell”), or “tricky” tasks (e.g., attention checks, memory checks). Second, and most relevant to the current discussion, some threads make suggestions on how to pass qualification screens. Using self-reported data, Chandler, Mueller, and Paolacci (2014) suggest that this behavior does occur but the extent of this distortion is unknown. Third, these forums provide a place for venting anger or frustration with requesters or other Turkers. Fourth, the forums encourage coworker friendship, which includes discussions of personal challenges that may or may not be related to completing MTurk tasks (Brawley and Pury 2016). The following table provides example quotes (some

Name (website)	Registered users ⁵	Open to the public? (Need for registration)
MTurk Forum (MTF) (http://www.mturkforum.com)	54,831	Yes. (No registration to view)
Hits Worth Turking For (HWTF) (https://www.reddit.com/r/HITsWorthTurkingFor)	35,626	Yes. (No registration to view)
MTurk Reddit (MTR) (https://www.reddit.com/r/mturk)	20,146	Yes. (No registration to view)
Turker Nation (TN) (http://www.turkernation.com)	17,891	No, this is a private site. (Requesters may sign up and receive limited access)
TurkerHub.com (TH) (https://turkerhub.com)	12,408 ⁶	Yes. (No registration to view)
Turk Opticon (TO) (https://turkopticon.ucsd.edu)	No user information published	Yes. (Need to register)
MTurk Crowd (MTC) (http://www.mturkcrowd.com/)	2,740	Yes. (No registration to view)

edited for clarity) that give a sense of how such MTurk communities operate.

MTurk community websites can thus generate problems for researchers by revealing experimental conditions, by undermining tests of respondent abilities or knowledge, or by enabling character misrepresentation that permits a person to enter a study under false pretense. It is important to note that such forums not only increase the risk of deception in studies but may also serve as a safeguard against such deception. For example, we conducted a 12-cent study with 736 Turkers who were asked to guess the number of gumballs in a jar with the ability to “win” a \$1 bonus if they guessed correctly. After each respondent made a guess, we revealed to the respondent the correct number of gumballs. We monitored whether the proportion of Turkers guessing the number correctly increased over time as well as the activity on MTurk forums to see if the correct answer was posted online. Indeed, shortly after we posted the study, a correct answer appeared briefly on HITsWorthTurkingFor (HWTF), notifying fellow Turkers of the response that would lead to the \$1 bonus. However, the post was criticized and taken down by the forum moderator within minutes (see the screenshot in web appendix C). As a result, relatively few people (3.8% of respondents) “guessed” the correct answer. Thus, while a small level of deception occurred, the moderator served to limit its impact by reinforcing norms of Turkers being reliable respondents.

A major function of the forum websites is to provide greater worker power. In particular, Turk Opticon (TO) was created to try to restore some balance of power between the workers and requesters. The TO platform allows Turkers to rate

requesters and comment on the HITs that requesters post based on four dimensions that workers care about: communicativity, generosity, fairness, and promptness. While separate from the MTurk platform, anyone may review the individual ratings from the TO site. Those with a Turker account may also load a browser script from Opticon that automatically generates and displays the requester’s aggregated Opticon scores while they browse for HITs on MTurk.

This drive for greater Turker control arose in part out of their perception that requesters are unfair because they have the ability to unreasonably reject or block Turkers. Through Amazon’s accept/reject functionality, requesters can reject a submission, and then not pay if a worker makes multiple attempts at a study, fails an attention check, does not submit the correct end-of-survey code, answers the survey too fast, makes a submission but never completes the study, or for any other reason. This rejection leads to immediate loss in income and negatively impacts the worker’s approval rating. Because requesters often set the requirement that Turkers have a particular approval rating (e.g., typically 95% or above), Turkers try to avoid anything that could hurt their rating. Further, a repeat offender may be blocked from all subsequent studies by that requester. Being blocked by several requesters can lead to the worker’s account being suspended and the worker being barred from completing any MTurk tasks. As a result, workers are highly sensitive to those actions that threaten their ability to work. The forums allow Turkers to quickly identify and disseminate requesters who commonly reject Turkers. While the forums restore some of the balance of power between requesters and Turkers, they may also discourage requesters from appropriately rejecting or blocking truly offending workers from their studies. Additionally, researchers sometimes do not reject or block offending Turkers because such processes require additional effort after the data collection has been completed. Instead, researchers are often motivated to quietly remove poor responses from their data. However, requesters who abstain from taking actions against deceptive

5 As of December 20, 2016.

6 TurkerHub.com was previously MTurk Grind (MTG; <http://www.mturkgrind.com/>), which had 12,408 registered users. User information for the newly created TurkerHub.com has not been published. However, daily views (by registered and nonregistered users) range from 8,984 to 46,213 (mean 18,855) during the second month of this forum’s inception.

Help passing a screener question

Must be in a romantic relationship to pass screener. (HWTF)

Attitude towards unpaid screeners at the beginning of a study

Unpaid screener. So sick of this crap, I wasted time reading the survey info. (MTC)

It's an annoyance. Requesters put up an unpaid screener, ask you enough questions to qualify as a paid survey, and then tell you that you aren't eligible. There really shouldn't be unpaid screeners - it gets abused and turned into mini-surveys. (HWTF)

I don't hate [unpaid screeners], as long as they're short, and not buggy. Ideally, they should also tell you that they have one, up-front. (MTR)

Help with avoiding attention and memory checks

Two attention checks. One requires you to recall a price, one requires you to write a word. (HWTF)

I always copy/paste whenever I see large blocks of text in case there is a memory check (this should not be considered cheating despite what others may say). (HWTF)

I was filling out a survey, failed an attention check, but I was able to retake the survey. Can requesters honestly see when Turkers do this? (MTF)

Help selecting HITS

Six and a half utterly unenjoyable minutes, but monetarily a HWTF.¹ (HWTF)

This one made me feel anti-social. Bubble hell warning. (HWTF)

Processing advice

I setup a macro using iMacro for each [option]. The attention checks are the same for a few days at a time so it comes down to how fast you can click one of the macro choices once you learn the pattern. (MTG)

Focus on speed

Finished a \$4 hit in less than 10 minutes so I decided to milk the timer. I've been rejected for going too fast but I'll milk the timer on a new requester who is over paying for hits, hoping that it will make them less likely to drop the pay. If they're paying \$4 and see people submitting hits in 8 minutes, the pay probably gets drastically reduced for their next hit. (MTG)

Socialization

Husband's birthday is on Tuesday and I'm like \$30 short of having enough to get him what I want to get him. Trying to get surveys done but not real hopeful of much getting approved over the weekend. (TH)

I work outside the home with 2 full jobs and Turk between. After a while it became easy to stay awake for a few days at a time without even getting that sleepy. Now I have to drug myself to even fall asleep. (TC)

Attitudes towards requesters

I am feeling like I need a mindless batch today. Very upset this morning to receive a rejection on a survey, emailed and asked why. . .they said I went too fast to have taken it seriously. I do all of the surveys carefully. . .I guess I need to let the clock play out. (TH)

I wish Amazon's improvements would include a Block Requester option. (MTC)

Terrible Requester. Seems to reject everybody. The goal of this character seems to be getting surveys done without actually paying. (TO)

Unintentionally revealing different stimuli conditions

Warnings about the same study:

Thread 1: 2 minutes writing

Thread 2: No writing in the version I did (HWTF)

¹ A HWTF ("HITs Worth Turking For") is any task that pays 10 cents or more per minute to complete. It is based on the actual time that a Turker took to complete the task and not the posted time by the researcher.

Turkers may be hurting the research community by not punishing these offenders.

Overall, the MTurk online forums help workers transform a difficult job of responding to studies into one that is more predictable, pleasant, and economically justifiable. In that way, forums benefit requesters by increasing the willingness of people to participate in research studies. Forums also encourage requesters to act in ways that support the joint system. In particular, the forums penalize requesters who pay a low hourly wage (Gleibs 2016), those who underreport the expected length of the study, those who annoy workers with unexpected or boring tasks, and those who block workers unjustifiably (Brawley and Pury 2016).

In effect, online MTurk communities serve as an informal labor union (Bederson and Quinn 2011), whereby

Turkers are able to lessen their efforts and improve their earnings through a collective system of notifying and warning fellow workers. Therefore, and as recently recommended by others (Cheung et al. 2016; Farrell, Grenier, and Leiby 2017), it is important for researchers to become familiar with these Turker communities and follow the chat-room discussions when a study is live. Doing so can help researchers evaluate how Turkers perceive the study, and whether their payment level is sufficient for the effort put into the study. It will also help researchers determine the extent to which screeners, attention checks, manipulations, or desired responses have been revealed to other Turkers.

POSSIBLE WAYS TO MINIMIZE CHARACTER MISREPRESENTATION

There are a number of ways to limit distortion from respondents who falsify their identities. We begin with a

⁷ A HWTF ("HITs Worth Turking For") is any task that pays 10 cents or more per minute to complete. It is based on the actual time that a Turker took to complete the task and not the posted time by the researcher.

number of solutions that are either infeasible or impractical, and then move to describe a version of a two-step process that can reduce, if not eliminate, the opportunity for deception.

Disguise Desired Screener Answers

Chandler and Paolacci (2017) have demonstrated that disguising a screener requirement reduces the amount of deception in MTurk studies. To make it more difficult for deception to occur, the screening questions should contain a number of items where it is hard to determine which responses will grant access to the study. However, it is often challenging to disguise a screener even if the researcher adds a list of possible options, because the respondent may still answer the questions in a way that maximizes her likelihood of qualifying for a study. For example, a respondent may claim product ownership for all (or of a larger number of) products to maximize the likelihood of passing the screen. Furthermore, Turkers often complain about being screened out of a study without being paid, without prior warning. Studies with disguised screeners are also susceptible to Turkers repeatedly taking the study (by clearing the cookies from their browsers) or to the leakage of screener criteria through the Turker communities.

Identify False Qualifiers after the Fact

Researchers commonly use attention checks or response time to screen respondents who are not sufficiently diligent (Peer, Vosgerau, and Acquisti 2014). Can similar approaches be effective for screening impostors *ex post*? Suppose one suspects that respondents have misrepresented their identity. Is there a way to adjust for it after the fact? Can one infer from responses to other questions or response style which respondents lied to get into a study compared to those who didn't? Unfortunately, the simple answer is no.

First, consider approval ratings. In our studies, we deliberately chose not to set an approval rating threshold so that we could assess the common requirement by researchers that Turkers should have a 95% approval rating to take their studies. The self-reported approval ratings gathered in our panel surveys had a mean approval rating of 99.1% with only 1% of our panelists under the 95% threshold, making it a difficult criterion to separate impostors from

those who answered honestly (Brawley and Pury 2016). Table 3 shows in the cell phone conjoint study that the average approval rating for impostors was 99.2% compared to a 99.1% approval rating for those who legitimately passed the screen. Indeed, across our five studies the average approval rating of impostors was not statistically different from that of honest respondents.

Table 3 also gives the results for traditional quality metrics. It shows that there is no statistically significant difference for failed attention and memory checks between those who deceived and those who honestly qualified in our cell phone study. Thus, including these in one's studies and either controlling for or eliminating those who fail these checks does not weed out impostors. Turkers, in general, are very good at detecting traditional attention checks (Farrell et al. 2017; Hauser and Schwarz 2015). There was also no difference in how much time one spent on the study between impostors and those who legitimately qualified. Finally, impostors and legitimately qualified respondents did not differ in regards to the conjoint fit statistic, RLH (Sawtooth Software 2013, 22). It appears that impostors are just as practiced and vigilant as honest Turkers.

We do find some demographic and psychographic differences between those who impersonate and those who are honest. There is preliminary evidence that extroverts ($p < .001$) and males ($p < .001$) on MTurk have a higher propensity to impersonate, but it would certainly not be desirable to remove everyone who fits these characteristics from a research study.

Pay All Respondents without Screening

We demonstrate that misrepresentation occurs rarely if there is no benefit from doing so. Therefore, if one is interested in a select group for pragmatic or theoretical reasons, a feasible solution is to simply collect information from everyone and statistically control for, or remove, undesired respondents from subsequent analyses. That strategy requires payment to unneeded respondents but has the advantage of providing information about the effect of individual differences. This approach is particularly attractive if the base rate of the screened population is relatively high. However, if the base-rate proportion of the screened population is low (e.g., people suffering from a particular disease), this

TABLE 3
CONJOINT STUDY QUALITY COMPARISON

Quality check	Impostors ($n = 35$)	Respondents who satisfied the screen ($n = 106$)	Significance
Failed attention check	0.0%	4.7%	$ z = 1.362; p = .173$
Failed memory check	11.4%	6.6%	$ z = .492; p = .622$
Approval rating	99.2%	99.1%	$ t = .464; p = .643$
Total time on study (minutes)	5.61	5.70	$ t = .100; p = .921$
Conjoint fit (RLH)	0.74	0.77	$ t = 1.136; p = .258$

approach can be prohibitively expensive. Still, one can limit wasted participants by moving respondents with undesired characteristics into other studies where those characteristics are desired. In a medical study, for example, those respondents 40 and over could take the lung cancer study, while those under 40 could take the shoulder dislocation study.

Use a Commercial Panel to Deliver Prescreened Respondents

Companies like Qualtrics and SSI provide access to prescreened panelists. However, these vendors tend to cost orders of magnitude more than managing the process oneself. Typical fees in 2016 are \$20 per completed 15-minute study compared with \$2 on MTurk. The price charged is generally much higher for rare populations. There are emerging enterprises, such as TurkPrime (Litman, Robinson, and Abberbock 2016) and Prolific Academic (ProA), that allow screening for a lower fee. Thus, we can expect the cost per respondent to decrease. However, while these commercial companies claim confidence in their prescreening, they offer little external verification. We encourage researchers who use such services to monitor and validate the quality of the screening. It is important for these organizations to test their panels just as our two-stage process tested the MTurk workers.

RECOMMENDED TWO-STEP APPROACH

We believe that prescreening participants before the focal study is the best way to reduce the expense of a study and limit the number of impostors. We first explain a one-off approach within MTurk and then describe a way to create and manage a panel of qualified respondents across multiple studies or researchers administered by a behavioral lab.

Run a Short Paid Prescreen

Researchers can run a prescreen questionnaire to establish who will be appropriate for a subsequent test, perhaps involving a simple \$.10 survey with a few quick questions. As mentioned above, it is important that the prescreen not be part of the actual study. If the actual study is desirable because it is highly paid or interesting, it is likely that the desired qualification conditions will be posted on an MTurk forum or that Turkers will attempt to retake the study. Additionally, it is important that the screening question be masked by other questions. For example, if one looks for respondents above a particular age or that own a particular product, the researcher should ask a few demographic and multiple product ownership questions in the paid prescreening questionnaire.

Develop an Ongoing Panel

Researchers who conduct multiple studies or coordinated studies within a behavioral lab setting could gain substantially by building an ongoing panel similar to the

one that we used to test the extent of misrepresentation. Figure 2 provides a flowchart for creating and managing such a panel. The panel could begin, as in our studies, with general questions to define a number of critical screening variables. Because any panel will gradually lose members over time, it is useful to include categorization questions in all studies that build information for future studies and test respondent consistency with earlier ones. With such a panel, studies that need a targeted population would be made available only to prescreened panel members. Even so, we recommend that a consistency check in the focal research study be included. For example, in a study where only females are permitted, we recommend including a gender question in the demographic section as a way to check for consistency with the initial panel response.

However, it would also be useful to allow a relatively small number of nonpanel members to take open studies to gradually develop and replenish the panel with new participants. It is also helpful to test panel members in various ways. For example, Chandler and Paolacci (2017) asked whether respondents own a brand that does not exist, or if they have rare diseases or do unlikely activities. Asking questions about impossible activities or fictitious events can help identify opportunistic, long-term, consistent deceivers. Note, however, that such questions should be used with caution, as Turkers are likely to catch on, especially if the question can be factually verified (Goodman, Cryder, and Cheema 2013).

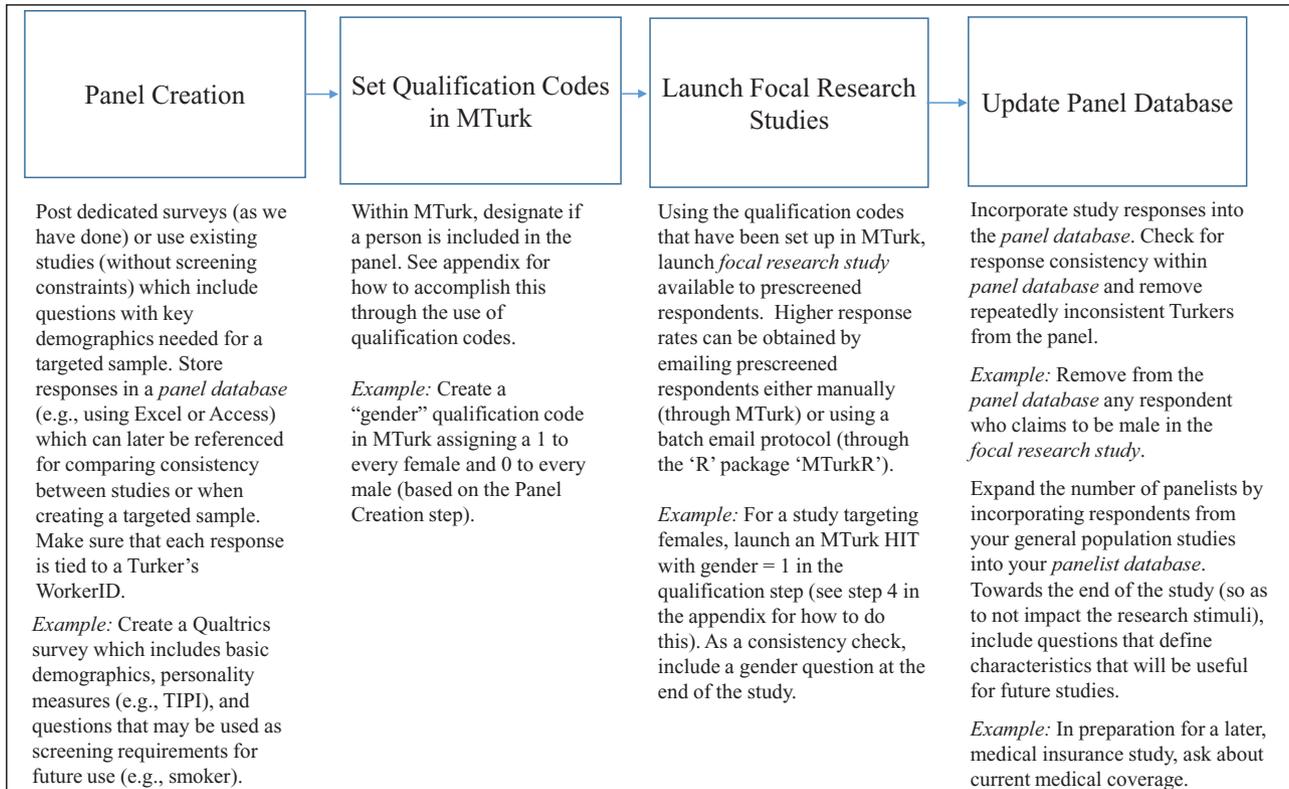
It is useful in setting up a panel to build a centralized repository for study responses. While a single researcher could easily manage such a data set in Excel, a robust system with more complex database management could emerge as part of a behavioral lab. In the ideal case, all MTurk studies would be managed through a central MTurk account that uses “qualification” codes to designate which Turkers would qualify based on prior responses. Web appendixes D and E explain the mechanics of using qualification codes for creating and managing a panel. The R package MTurkR is useful in creating and updating qualification codes once the panel size becomes sufficiently large (Leeper 2017). This package is also helpful for sending batch emails to notify prequalified respondents that they are eligible for new studies. In this way, a researcher or lab coordinator can manage an MTurk pool, similar in nature to a professional panel company or student participant pool, while benefiting from the relatively low cost of using MTurk.

DISCUSSION

There are four goals to this tutorial. First, we demonstrate the extent to which character misrepresentation occurs when Turkers are given the opportunity to do so. Deceivers, having gained access to a desired study, distort their identities and can generate unstable responses to later questions. Second, we provide evidence that MTurk workers

FIGURE 2

PANEL CREATION AND MANAGEMENT PROCESS



are very consistent when there is no motive to lie. Third, we explore the motivations and activities of Turkers as revealed by their comments on MTurk forums. We advocate and detail a two-step process where the first step is to identify appropriate respondents and the second is to target directly those who qualify. Finally, we recommend that this two-step process be incorporated within a larger panel management system.

The fact that the results of MTurk studies depend on how each study is introduced and managed within the system implies that more effort is needed to document how a study is implemented and how respondents are recruited. Scientific progress requires others to be able to replicate a study, and as a field, we need to move toward including the kinds of detail shown in the following table as part of the study reporting. Of course, not all of this information is needed for every study, but such detail is appropriate in a web appendix to help readers better understand and be able to replicate the work.

Perhaps the greatest lesson from recent work demonstrating the likelihood of deceit from Turkers is the need for constant vigilance on the part of researchers. Such vigilance requires a number of efforts, such as including validation tests that ask the same question in different ways and

checking for consistency. Unlike categorical and substantial lies, such softer inconsistency only suggests a heightened probability of deceit or undesired sloppiness. The question then arises of the appropriate reaction on the part of a researcher who suspects that a Turker is behaving irresponsibly. One response is to reject the Turker's submission, an action that will reduce the Turker's approval rating. Requesters may also block the Turker from taking future studies. Both solutions are quite effective in penalizing the individual Turker but can result in an unfair penalty for an honest mistake or inconsistency, as well as negative reactions against the researcher if the incident is disseminated within the Turker communities. An alternative response is to remove the respondent from the panel, which eliminates the possibility that the respondent will contaminate future studies. Such actions are better for both the individual Turker and researcher in the short term. However, the formal action of rejecting the submission or blocking the respondent from taking future studies provides a greater benefit to the entire research community, which gains from holding our participants accountable for honest and dishonest responses. We encourage researchers to contribute to the

Characteristics of the study as posted on MTurk

How were Turkers recruited to take the study (i.e., wording of the HIT description)?

The expected time to completion

Notification if there is an unpaid screener

Screening process

Was screening part of the focal study (unpaid) or completed as part of a previous study (paid)?

The exact wording of the screening question(s) and which options led to being screened out

Percent of respondents attempting to start the study but failing the screener

Completion history

Average and standard deviation of completion time

Date and time survey opened and closed

The number of times the study was posted/reposted (i.e., study launched in micro-batches)

Attrition: percent of respondents quitting before the end of the study by condition

Avoiding multiple responses

Was a back button allowed?

Was “Prevent Ballot Box Stuffing” implemented?

Micro-batches (if applicable): how were multiple responses prevented or screened out?⁸

Sample cleansing

Percent of respondents dropped due to failed attention, memory, consistency, or speed checks

Were multiple attempts by the same respondents removed? If so, how many were removed?

Vigilance

Monitoring of specific MTurk communities

Reporting any MTurk community discussions that could be relevant to the research results

community by flagging poor-quality Turkers, but because such actions will have a direct effect on a Turker’s source of income, we recommend doing so only when the dishonesty is clear and disruptive to scientific progress.

Finally, we build on [Goodman and Paolacci \(2017\)](#)’s tutorial in urging consumer behavior researchers who use MTurk workers for their studies to better understand these participants and treat them as important contributors to their research ([Gleibs 2016](#)). Thus, it is important that HIT descriptions help respondents find topics that they can manage well and even enjoy ([Brawley and Pury 2016](#)). Researchers also need to avoid the negative surprises from hidden tests that lead to frustration or anger. Ironically, strong positive surprises can also be distorting if they encourage respondents to misrepresent themselves to gain access. As a long-run proposition, we find that building a stable but continuously refined MTurk panel improves both parties. The MTurk workers gain from steadier and more predictable work from a regular source, while the researchers gain from a loyal, dependable panel about which much is known before the study begins.

DATA COLLECTION INFORMATION

The first author collected the data for the eight panel surveys (leading to the panel creation), the five deception tests, and the gumball study on Amazon Mechanical Turk from June 2015 to February 2016. Funds to collect this

data were provided from the institutional research budgets of all three authors. Analysis of the data was completed by the first author with oversight from the second and third authors from February 2016 to February 2017.

APPENDIX USING QUALIFICATION CODES TO CREATE AN MTURK PANEL

This appendix is primarily focused on creating and using Qualification codes within MTurk for the purposes of managing a participant pool on MTurk. Qualifications are particularly useful in accomplishing the following:

- Designating your panelist: indicating which workers (“Turkers”) are to be included in your panel (procedure described here).
- Prequalifying participants for a study: indicating if a participant (after taking a prequalifying survey) meets certain requirements (e.g., respondent is female) for taking a future study (see web appendix D for procedure).
- Removing participants from your panel: this is a way to “soft block” participants from taking future studies (see web appendix D for procedure).

Creating a panel using qualification codes within MTurk involves the following four steps:

1. Create a new qualification type (to be used to designate whether or not someone is in your panel).
2. Download the Worker file and assign Turkers to your panel.
3. Upload the updated Worker file (which include your panelist designations).
4. Include your new panel designations as a criterion when launching a new MTurk study.

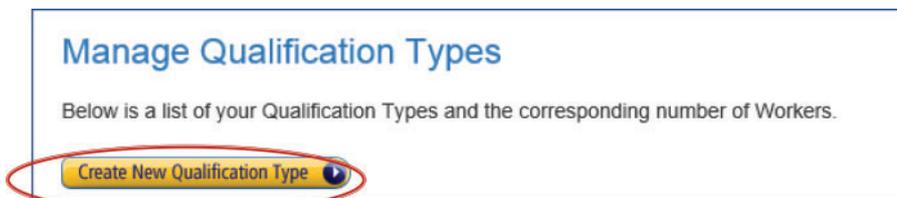
⁸ Micro-batches are when a researcher launches the same study multiple times in order to achieve the desired sample size. Each time the study is launched, the MTurk platform places it at the top of the queue of HITs, which may result in faster completion times.

Step 1: Create Qualification Type

To form a panel within MTurk, click on the Manage tab and then Qualification Types within your MTurk Requester account.



Click on the Create New Qualification Type button.



For your qualification type, name your panel by entering a label under the Friendly Name field. As it is required by MTurk, provide a description. *Note: Turkers will be able to view your name and description (which is required) so it is advised that you keep your qualification names and descriptions general, but specific enough for you to remember why you are using these.* We labeled our qualification name “qual,” which is short and generic.

Create New Qualification Type ×

This will create a new Qualification Type that can be assigned to Workers.

Friendly Name:

Description:

When the new qualification type has been created, you should be able to view it in the Manage Qualification Types table within the MTurk interface. It may take a few minutes for the system to update, and you will need to refresh the page to view. When your new qualification type has been created, there will be a 0 in the “Workers who have this Qualification” column, as workers have not yet been added to your panel.

Manage Qualification Types

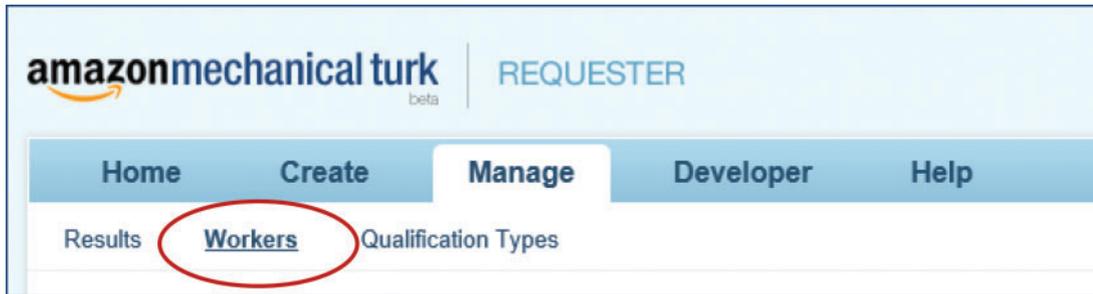
Below is a list of your Qualification Types and the corresponding number of Workers.



Qualification Types				
	Name ▼	ID	Workers who have this Qualification	Description
⊗	qual	3Y9DA377APTUXN8YBHSZHJZW5QGAF	0	included in panel

Step 2: Download the Workers File and Tag Each Participant (by Worker ID)

To add participants to your panel, download your global MTurk Workers file. To do so, click on Workers under the Manage tab.



Here you will find a list of all of the Turkers who have ever completed a HIT for you. For each Turker, in the first column is the Worker ID, and in the second column are the number of HITs that they have completed and the number that you have approved. For example, the fourth Turker on the list below has completed eight of our studies, and we have approved all eight of his or her submissions (as reflected in the lifetime approval rate). This 100% approval rating is just for *our* studies and does not incorporate the approval ratings from other researchers (i.e., Requesters).

Next, click on the Download CSV button to export this table.

The screenshot shows the 'Manage Workers' page. At the top, there is a title 'Manage Workers' and a description: 'The Workers who have completed work for you are listed below. Select a Worker ID to bonus, block, unblock, assign a Qualification, or revoke a Qualification. To block, unblock, or change Qualification settings for multiple Workers, select Download CSV. Select Customize View to change which Qualification Types are displayed in the table below.' Below this, there is a 'Customize View' button and a 'Download CSV' button circled in red. There is also an 'Upload CSV' button. Below the buttons, there is a 'Show my Workers by:' section with 'Lifetime', 'Last 30 days', and 'Last 7 days' options. To the right, there is a pagination control: '← Previous 1 2 3 4 5 6 7 8 9 ... 195 196 Next →'. Below this is a table with the following data:

Worker ID ▲	Lifetime Approval Rate for Your HITs	Qual: qual...	Block Status
A0137396XUHSYEXAMPLE	100% (1/1)		Never Blocked
A03045182ACQPEXAMPLE	100% (1/1)		Never Blocked
A031662732J1CEXAMPLE	100% (1/1)		Never Blocked
A0501481M8MAEXAMPLE	100% (8/8)		Never Blocked
A050383932MZWEEXAMPLE	100% (3/3)		Never Blocked

This .csv file includes a list of every worker who has ever completed a study for you. In addition to the lifetime stats (pertaining to your studies) for each individual, you will find two columns for each qualification type that you have created. The columns are automatically named with the following convention: CURRENT-Friendly Name and UPDATE-Friendly Name, where Friendly Name, refers to the name that you chose to call your panel. In our example, our Friendly Name is “qual” so the two columns associated with our panel are CURRENT-qual and UPDATE-qual.

A	C	D	E	L	M
Worker ID	Number of	Number of	Your Lifeti	CURRENT-qual	UPDATE-qual
A1RJ2LOEXAMPLE	1	1	100.00%		
A8DRC9EXAMPLE	1	1	100.00%		
76C9EXAMPLE23	1	1	100.00%		
8UHC9EXAMPLE2	1	1	100.00%		

To add a worker to your panel, assign a numerical code (anywhere from 0 to 100) in the UPDATE column. We use the following convention when creating a panel: 1 to anyone in our panel and blank for everyone else.⁹ In our example .csv file, we have entered a 1 in the UPDATE-qual column for the following Workers IDs: A1RJ2LOEXAMPLE, A8DRC9EXAMPLE, and 8UHC9EXAMPLE2. Thus, when this procedure is complete, these three workers will be included in our newly created panel.

A	C	D	E	L	M
Worker ID	Number of	Number of	Your Lifeti	CURRENT-qual	UPDATE-qual
A1RJ2LOEXAMPLE	1	1	100.00%		1
A8DRC9EXAMPLE	1	1	100.00%		1
76C9EXAMPLE23	1	1	100.00%		
8UHC9EXAMPLE2	1	1	100.00%		1

When you are finished with revising this Worker file, save it as a .csv file to be used in the next step.

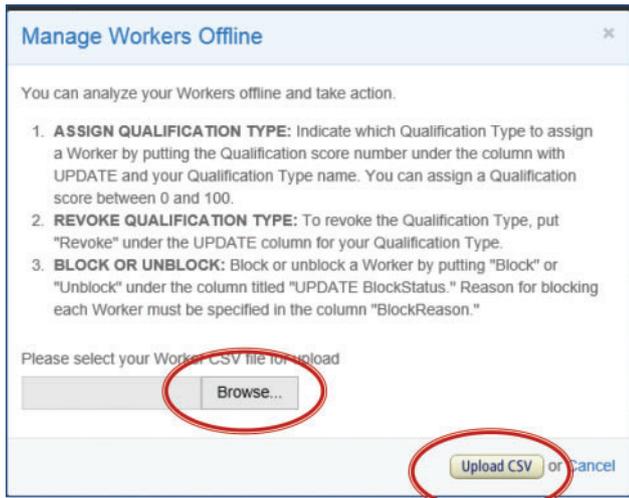
Step 3: Upload the Updated Worker File

To officially create this panel to be used within MTurk, you need to upload the revised .csv Worker file. To do so, click on the Upload CSV button (under the Manage Workers tab in MTurk).

The screenshot shows the 'Manage Workers' interface in MTurk. At the top, there is a 'Customize View' button and a 'Download CSV' button. The 'Upload CSV' button is circled in red. Below these buttons, there are filters for 'Show my Workers by: Lifetime, Last 30 days, Last 7 days' and a pagination control 'Previous 1 2 3 4 5 6 7 8 9 ... 195 196 Next'. The main content area is a table with columns for 'Worker ID', 'Lifetime Approval Rate for Your HITs', 'Qual: qual...', and 'Block Status'.

⁹ We leave the space blank for Turkers that we do not have enough information about to discern if they should be in our panel. If we know at this point that someone should not be in our panel (e.g., Turkers that have demonstrated inconsistency or deception in the past), we would assign a 0 to the qual code of these individuals.

Next, select your .csv file (click Browse) and click Upload CSV. *Note: Excel files do not work within the MTurk environment. If you have your updates saved in an Excel file, convert it to a .csv file before uploading.*



Throughout this process, you may have noticed that you have an option to block specific Turkers from ever taking future studies (in the Block Status column). We recommend against using this feature, as in our experience it leads to emails from Turkers concerned about their MTurk accounts being revoked. Qualification codes are a far more effective way to limit who is allowed to take part in your studies.

Once you have uploaded your revised Worker file (.csv), you have created your panel. You will see on the screen which workers are included and which ones are not. In our example, there is a qualification named "qual," and some Turkers (each having a unique Worker ID) have been assigned the value of 1.

Manage Workers

The Workers who have completed work for you are listed below. Select a Worker ID to bonus, block, unblock, assign a Qualification, or revoke a Qualification. To block, unblock, or change Qualification settings for multiple Workers, select Download CSV. Select Customize View to change which Qualification Types are displayed in the table below.

Customize View **Download CSV** **Upload CSV**

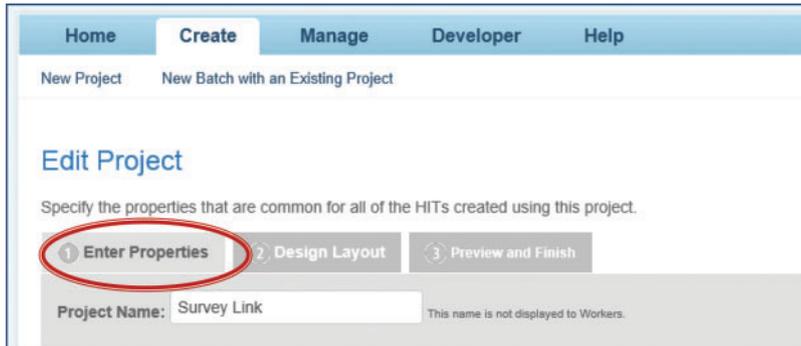
Show my Workers by: **Lifetime** Last 30 days Last 7 days

← Previous 1 2 3 4 5 6 7 8 9 ... 195 196 Next →

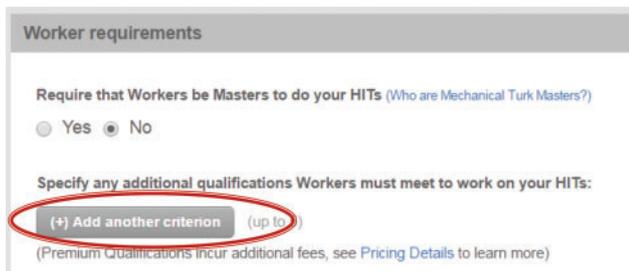
Workers			
Worker ID ▲	Lifetime Approval Rate for Your HITs	Qual: qual...	Block Status
A01373EXAMPLE	100% (1/1)		Never Blocked
A0304518EXAMPLE	100% (1/1)		Never Blocked
A031662732EXAMPLE	100% (1/1)		Never Blocked
A1RJ2LOEXAMPLE	100% (8/8)	1	Never Blocked
A8DRC9EXAMPLE	100% (3/3)	1	Never Blocked

Step 4: Using Your Panel for Future Studies

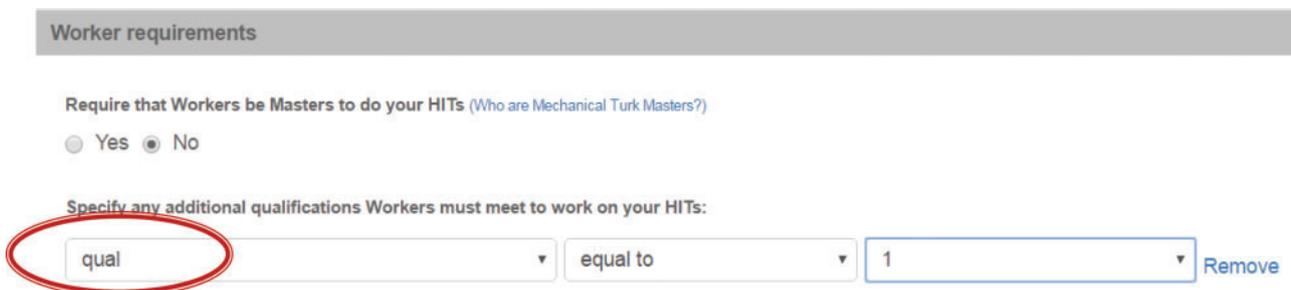
Assume that you want to make your next study available only to your panelists. When creating your HIT, click on the Enter Properties tab.



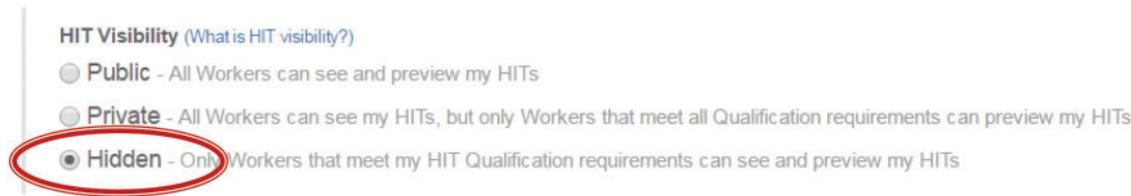
Scroll down to the “Worker requirements” section and click “(+ Add another criterion” button.



Scroll down to the “Qualification Types you have created” section within the drop-down menu and select your panel name (this is the Friendly Name from earlier). In our example, “qual” is selected and set “equal to” the value of 1, indicating that only panelists are eligible to take part in our studies.



In the HIT Visibility section, make sure that Hidden has been checked, indicating that only your panelists can view and take part in your study. Otherwise, you may receive email requests from nonpanelists requesting that they be added to your panel. If this is undesirable, make sure that Hidden is checked.



Then continue to post your new HIT as usual. Note, to improve the response rate, you may want to notify Turkers of the new study that you have posted. Unfortunately, there is no easy way to do this within the MTurk platform. You would need to click on each Worker ID and manually send a personal email to each Turker who qualifies. The R package MTurkR does allow for batch notifications. See web appendix E for example code for sending out batch notifications.

REFERENCES

- Barriga, Alvaro Q., Elizabeth M. Morrison, Albert K. Liau, and John C. Gibbs (2001), "Moral Cognition: Explaining the Gender Difference in Antisocial Behavior," *Merrill-Palmer Quarterly*, 47 (4), 532–62.
- Bederson, Benjamin B. and Alexander J. Quinn (2011), "Web Workers Unite! Addressing Challenges of Online Laborers," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, New York: ACM, 97–106.
- Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012), "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Political Analysis*, 20 (3), 351–68.
- Brawley, Alice M. and Cynthia L. S. Pury (2016), "Work Experiences on MTurk: Job Satisfaction, Turnover, and Information Sharing," *Computers in Human Behavior*, 54, 531–46.
- Burke, Daniel (2016), "Pope Suggests Trump 'Is Not Christian,'" CNN Politics, February 18, 2016, <http://www.cnn.com/2016/02/18/politics/pope-francis-trump-christian-wall/>.
- Chandler, Jesse, Pam Mueller, and Gabriele Paolacci (2014), "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers," *Behavior Research Methods*, 46 (1), 112–30.
- Chandler, Jesse and Gabriele Paolacci (2017), "Lie for a Dime: When Most Prescreening Responses Are Honest but Most Study Participants Are Imposters," *Society Psychological and Personality Science*, forthcoming.
- Chandler, Jesse, Gabriele Paolacci, Pam Mueller, Eyal Peer, and Kate A. Ratliff (2015), "Using Nonnaïve Participants Can Reduce Effect Sizes," *Psychological Science*, 26 (7), 1131–9.
- Cheung, Janelle H., Deanna K. Burns, Robert R. Sinclair, and Michael Sliter (2016), "Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations," *Journal of Business and Psychology*, 1–15.
- Crowne, Douglas P. and David Marlowe (1960), "A New Scale of Social Desirability Independent of Psychopathology," *Journal of Consulting Psychology*, 24 (4), 349–54.
- Dholakia, Utpal (2015), "My Experience as an Amazon Mechanical Turk (MTurk) Worker," July 20, <https://www.linkedin.com/pulse/my-experience-amazon-mechanical-turk-mturk-worker-utpal-dholakia>.
- Farrell, Anne M., Jonathan H. Grenier, and Justin Leiby (2017), "Scoundrels or Stars? Theory and Evidence on the Quality of Workers in Online Labor Markets," *Accounting Review*, 92 (1), 92–114.
- Gleibs, Ilka H. (2016), "Are All 'Research Fields' Equal? Rethinking Practice for the Use of Data from Crowdsourcing Market Places," *Behavior Research Methods*, 1–10.
- Goodman, Joseph K., Cynthia E. Cryder, and Amar Cheema (2013), "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples," *Journal of Behavioral Decision Making*, 26 (3), 213–24.
- Goodman, Joseph K. and Gabriele Paolacci (2017), "Crowdsumers Take Over: Towards Valid Crowdsourcing of Consumer Research," *Journal of Consumer Research*, forthcoming.
- Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann (2003), "A Very Brief Measure of the Big-Five Personality Domains," *Journal of Research in Personality*, 37 (6), 504–28.
- Gould, Stephen J. (1988), "Consumer Attitudes Toward Health and Health Care: A Differential Perspective," *Journal of Consumer Affairs*, 22 (1), 96–118.
- Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto (2011), "Mapping the Moral Domain," *Journal of Personality and Social Psychology*, 101 (2), 366–85.
- Haidt, Jonathan, Jesse Graham, and Craig Joseph (2009), "Above and Below Left–Right: Ideological Narratives and Moral Foundations," *Psychological Inquiry*, 20 (2–3), 110–9.
- Hauser, David J. and Norbert Schwarz (2016), "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks Than Do Subject Pool Participants," *Behavior Research Methods* 48 (1), 400–7.
- John, Oliver P. and Sanjay Srivastava (1999), "The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives," *Handbook of Personality: Theory and Research*, 2, 102–38.

- Leeper, Thomas J. (2017), "MTurkR: R Client for the MTurk Requester API, 2017," R Client for MTurk Requester API, <https://cran.r-project.org/web/packages/MTurkR/MTurkR.pdf>. R package version 0.8.0.
- Leitch, Will (2004), "Group Thinker," *New York Magazine*, June 21, <http://nymag.com/nymetro/shopping/features/9299/#comments>.
- Litman, Leib, Johnathan Robinson, and Tzvi Abberbock (2016), "TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences," *Behavior Research Methods*, forthcoming.
- Martin, David, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta (2014), "Being a Turker," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, New York: ACM, 224–35.
- Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis (2010), "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, 5 (5), 411–9.
- Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti (2014), "Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk," *Behavior Research Methods*, 46, 1023–31.
- Richins, Marsha L (2004), "The Material Values Scale: Measurement Properties and Development of a Short Form," *Journal of Consumer Research*, 31 (1), 209–19.
- Ross, Joel, Lilly Irani, M. Silberman, Andrew Zaldivar, and Bill Tomlinson (2010), "Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, New York: ACM, 2863–72.
- Sawtooth Software (2013), "The CBC System for Choice-Based Conjoint Analysis," Technical Paper Series, <https://www.sawtoothsoftware.com/download/techpap/cbctech.pdf>.
- Sharpe, Kathryn, Richard Staelin, and Joel Huber (2008), "Using Extremeness Aversion to Fight Obesity: Policy Implications of Context Dependent Demand," *Journal of Consumer Research*, 35 (3), 406–22.
- Tong, Betty C., Joel Huber, Deborah D. Ascheim, John Puskas, T. Bruce Ferguson Jr., Eugene Blackstone, and Peter K. Smith (2012) "Weighting Composite Endpoints in Clinical Trials: Essential Evidence from the Heart Team," *Annals of Thoracic Surgery*, 94 (6), 1908–13.
- Wessling, Kathryn Sharpe, Oded Netzer, and Joel Huber (2016), "Customer Response to Within-Chain Price Hikes," working paper.
- Zhou, Haotian and Ayelet Fishbach (2016), "The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions," *Journal of Personality and Social Psychology*, 111 (4), 493–504.