Computers in Human Behavior 83 (2018) 243–253

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Full length article

Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce

Irene P. Kan*, Anna B. Drummey

Department of Psychological and Brain Sciences, Villanova University, Pennsylvania, USA

ARTICLE INFO

Article history:
Received 23 June 2017
Received in revised form
4 February 2018
Accepted 6 February 2018
Available online 7 February 2018

Keywords:
Mechanical Turk
Online data collection
Demographic deception

ABSTRACT

Amazon's Mechanical Turk (MTurk) is fast becoming the most popular online research platform, and as such, it is crucial for researchers to recognize its advantages and shortcomings. Here, we focused on the issue of worker deception and examined the downstream consequences of demographic misrepresentation in MTurk. In two studies, we asked: “Are we testing who we think we are testing?” and “Does demographic deception ultimately have an impact on data quality?” We found that in the presence of explicit eligibility requirements, an alarmingly high proportion of our samples misrepresented themselves in order to qualify for the studies (55.8% in Study 1 and 21.8% in Study 2). We also found that the nature of the downstream consequences of demographic deception varied across studies. Specifically, the scope of the impact may rest with the relationship between the demographic variable of interest and the outcome measure. In sum, the impact of demographic deception on data quality is multi-faceted, and a fruitful avenue of future research is to identify additional motivating factors that may underlie such deception.

1. Introduction

Online data collection has permeated and transformed many research enterprises over the last twenty years, ranging from consumer research to Political Science to Psychology (Wolfe, 2017). Over the last decade, its popularity has risen even further, aided by the advent of several online behavioral research platforms, such as Amazon's Mechanical Turk (MTurk; Bohannon, 2011; Buhrmester, Kwang, & Gosling, 2011; Chandler, Mueller, & Paolacci, 2014; Pontin, 2007), Amazon's TurkPrime (Litman, Robinson, & Abberbock, 2016), and open-source framework like psiTurk (Gureckis et al., 2016). In a recent review, Chandler and Shapiro (2016) estimated that, between 2006 and 2014, approximately 15,000 published papers used MTurk as a source of data collection. In another review, Stewart, Chandler, and Paolacci (2017) projected that, in a few years, close to half of the articles published in cognitive science will involve samples from online crowdsourcing platforms. Indeed, the online labor force, and MTurk in particular, has become so popular that some researchers have called it “the new fruit fly for applied psychological research” (Highhouse & Zhang, 2015).

As data collection from the online workforce continues to gain momentum, there is a corresponding rise of investigations that assess the potential advantages and disadvantages of online data collection. We contribute to this growing literature by directing our empirical examination to a previously under-explored risk, namely the downstream consequences of worker deception in the MTurk environment (see Fig. 1 for the anatomy of an MTurk study). We, and many others, have focused on MTurk in these methodological investigations because it is the most popular, but it should be noted that many of the concerns identified below are not unique to MTurk. In fact, they likely generalize to other online data collection mechanisms, including professional panel studies (e.g., Kees, Berry, Burton, & Sheehan, 2017; but see Peer, Brandimarte, Samat, & Acquisti, 2017). We will begin our discussion with a brief review of the benefits and limitations associated with online data collection.

* Corresponding author. Department of Psychological and Brain Sciences, Villanova University, 800 E. Lancaster Avenue Villanova, PA 19085, USA.
E-mail address: irene.kan@villanova.edu (I.P. Kan).

https://doi.org/10.1016/j.chb.2018.02.005
0747-5632/© 2018 Elsevier Ltd. All rights reserved.

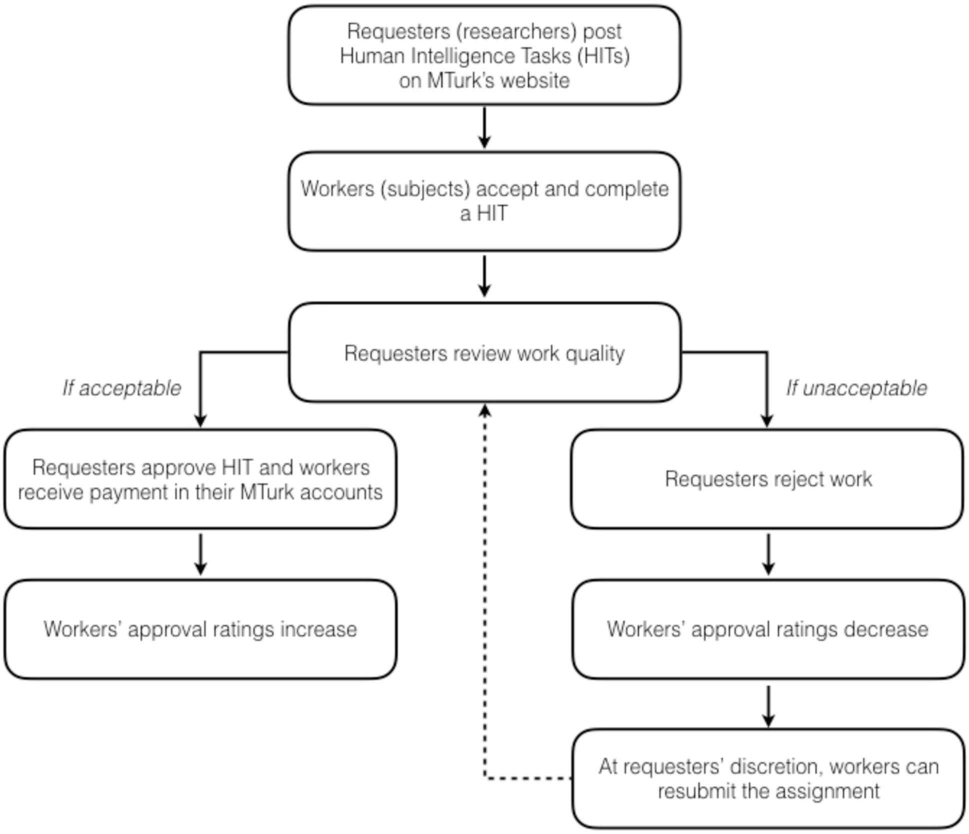
Requesters (researchers) post Human Intelligence Tasks (HITs) on MTurk's website
Workers (subjects) accept and complete a HIT
Requesters review work quality
If acceptable
If unacceptable
Requesters approve HIT and workers receive payment in their MTurk accounts
Requesters reject work
Workers' approval ratings increase
Workers' approval ratings decrease
At requesters' discretion, workers can resubmit the assignment

Fig. 1. A flowchart outlining the basic steps of an MTurk study.

1.1. Potential advantages of MTurk

The most commonly identified advantages of MTurk/TurkPrime[1] are efficiency, cost-effectiveness, relative anonymity, and diversity (Bohannon, 2011; Buhrmester et al., 2011; Casler, Bickel, & Hackett, 2013; Litman et al., 2016). For instance, traditional laboratory studies that would take weeks to complete and cost hundreds of dollars may take only several hours on MTurk, at a fraction of the cost. In addition, the relative perceived anonymity provided by an online platform presumably encourages respondents to be more candid (see Lease et al., 2013 for a discussion on MTurk worker anonymity; Paolacci, Chandler, & Ipeirotis, 2010). Finally, the diversity of the MTurk workforce could also mitigate the concern of drawing broad conclusions about human behavior based on the typically homogeneous samples in most Psychology studies (Arnett, 2008; Henrich, Heine, & Norenzayan, 2010a, 2010b). Indeed, Amazon's promise of access to more than 500,000 workers from 190 countries is enticing (Buhrmester et al., 2011; Casler et al., 2013), as it offers far greater age, cultural and socioeconomic diversity than most samples accessible by researchers (see Stewart et al., 2015 for an effective sample size estimate).

In addition to the above advantages, the growing evidence of comparable results between MTurk and traditional laboratory studies is another driving force behind the proliferation of MTurk studies. Across a variety of tasks in different domains, researchers found that patterns of findings are often similar across the two sources of data and concluded that MTurk is a viable alternative to the more traditional method of in-person testing (e.g., Bates & Lanza, 2013; Casler et al., 2013; Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Horton, Rand, & Zeckhauser, 2011; Kees et al., 2017). In fact, some researchers have reported that MTurk respondents are actually more attentive to task instructions than face-to-face subjects (e.g., Hauser & Schwarz, 2016; Kees et al., 2017).

1.2. Potential concerns of using MTurk and other online labor force

Despite the benefits summarized above, some researchers have raised serious concerns about methodological rigor and data quality associated with this data source (e.g., Chandler et al., 2014, 2015; Crump, McDonnell, & Gureckis, 2013; Deneme, 2016; Goodman & Paolacci, 2017; Harms & DeSimone, 2015; Rouse, 2015). Some of these issues include: (a) the potential impact non-naïve subjects may have on the results, (b) lack of control over testing environment, and (c) potential deception by respondents. Since many MTurk respondents are quite prolific and participate in many studies (e.g., Chandler et al., 2014, 2015; Peer et al., 2017; Stewart et al., 2017), some of them may not be naïve to common research material and procedures, and such prior exposure may contaminate and skew the results (e.g., reduction of effect sizes, Chandler et al., 2015). Moreover, it has been suggested that testing conditions can also be quite variable. Although most workers reported completing the tasks in a relatively distraction-free environment, many respondents admitted to engaging in other activities while completing Human Intelligence Tasks (HITs), and

[1] For the sake of simplicity, we will use the term "MTurk" throughout the manuscript when referring to the MTurk and TurkPrime workforce collectively. For a treatment of the differences between MTurk and TurkPrime, see Litman et al. (2016).

these other activities include watching television, listening to music, and instant messaging (Chandler et al., 2014; see Necka, Cacioppo, Norman, & Cacioppo, 2016 for a comparison of distracting behaviors across different samples). In this paper, we will focus on the last of these concerns: potential deception by MTurk respondents.

1.2.1. *Worker deception and its downstream consequences*

Recent studies have identified different ways in which online workers may deceive researchers. Here, we focus on the case of demographic misrepresentation, where workers lie about their demographic characteristics in order to gain entry into studies. It has been reported that such deception is most likely to occur when explicit screening criteria are in place (Chandler & Paolacci, 2017; Wessling, Huber, & Netzer, 2017). Given that most workers have strong monetary and internally-motivated (e.g., enjoyment) reasons to maximize task completions (Buhrmester et al., 2011), it is not surprising that they are more likely to lie when their eligibility to take part in a study is threatened. What is less clear, however, is the magnitude and consequences of this type of deception.

In a recent paper, Chandler and Paolacci (2017) compared deception rates under conditions of presence and absence of explicit eligibility requirement. They found that an explicit screening criterion led to a 45.3% deception rate (self-reported lesbians/gays/bisexuals who are actually heterosexuals), compared to a deception rate of 3.8% when no inclusion criterion was stated (Study 2). The authors asserted that Turkers are generally honest, except when incentivized to lie (see also Chandler & Shapiro, 2016; Rand, 2012; Wessling et al., 2017). In another study, the authors found that the percentage of respondents who claimed to have a child with autism jumped from 4.3% to 7.8% when "having a child with autism" was a criterion for eligibility for participation in future studies. At first glance, this increase may not seem consequential. However, the authors underscored the magnitude of the effect of this deception by stating that, "due to the rarity of autism, 45% of the self-identified eligible participants in the explicit prescreening condition are probably deceptive." Thus, if researchers were to attempt to draw conclusions about autism from a sample like this, almost half the data would be derived from parents who do not have children with autism! Importantly, this pattern of deception also extended to several other demographic characteristics. To summarize, these data highlight the potentially harmful consequences of utilizing explicit prescreening criteria for targeting a specific sample in the MTurk environment. However, the nature of the potential downstream consequences of such misrepresentations remains under-specified in these studies.

To fill this gap, Wessling et al. (2017) investigated the ways in which misrepresentations may threaten data quality. First, Wessling and colleagues replicated Chandler and Paolacci's (2017) finding that when explicit prescreening is absent, few people lie about their characteristics. However, when respondents do lie, there exist significant downstream consequences. In one of the studies, they compared survey responses from honest participants (target group 50 and older) and imposters (self-labeled as 50 and over but with average actual age of 33). They found that the imposters reported a higher frequency of fiber supplement usage than the honest participants. The researchers postulated that given the imposters' lack of personal knowledge, they were likely relying on stereotypes to answer questions, resulting in misleading data that would lead to flawed generalizations. Wessling and colleagues found a similar pattern of results in another study, where imposters (self-report female who are actually male) responded with stereotypical female answers to survey questions about cell phone case design and color choices (i.e., preference for ultra slim design and pink). Interestingly, the fact that imposters relied on stereotypes in their responding suggests that they might have been trying to appear authentic by maintaining their false identities throughout the task. However, only one critical question was used to assess these attempts. Thus, it remains unclear whether imposters could maintain their false identities, albeit fruitlessly, when the task requires more time and effort.

In sum, these recent findings of demographic misrepresentations suggest that the common (and often necessary) practice of employing explicit prescreening criteria will likely encourage purposeful demographic deception by online respondents. As such, it is essential to further examine this issue and to gain a better understanding of how worker deception may threaten data quality and reliability.

1.3. *Current studies — characterizing the downstream consequences of demographic deception*

We ask whether the deceitful act of demographic misrepresentation would necessarily translate to poor data quality. In other words, do imposters lie just to gain entry into a study, or do their duplicities extend to other aspects of the task? Across two studies, we examined the downstream consequences of demographic deception in the context of a genetic condition (Study 1) and a complex psychological construct (Study 2).

2. Study 1

To assess demographic misrepresentation, we employed a procedure where deception can be detected based on the pattern of responses on a behavioral task. Specifically, we recruited color-blind individuals on MTurk and administered parts of the Ishihara Test, a color vision test that has been in use since 1917 (Ishihara, 1972). Briefly, each test stimulus, called a plate, is composed of dots that vary in size and color, and the dots are arranged in such a way that a number or shape would be visible to individuals with normal color vision but not visible to those with red-green color-blindness. It should be noted, however, that we are *not* interested in subjects' responses to these pseudo-isochromatic plates. Given the variability in lighting conditions and computer screen settings, color vision assessment with an online version of the task in an uncontrolled environment would be quite unreliable. As such, it is best to think of the pseudo-isochromatic plates as filler trials. Of critical interest to our study are the workers' responses to plates that are intentionally designed such that numbers would be visible to *all* individuals, regardless of the integrity of their color vision. We will refer to these as "critical trials." We hypothesized that honest respondents who are in fact color-blind would report seeing numbers on these critical trials, whereas imposters would report seeing only dots on the critical trials. The assumption here is that imposters would rely on the common misconception that color-blind individuals see the world in shades of gray (Wisconsin Department of Health Services, 2016), and they would resort to this knowledge as they attempt to maintain their façade. Following this logic, the erroneous expectation is that when confronted with a display of colored dots, a color-blind individual would see gray dots and would have difficulties perceiving the embedded number. Unbeknownst to the imposters, the critical trials are designed in such a way that the specific color combinations do not preclude color-blind individuals from seeing the embedded numbers. In sum, an imposter is indexed by the combination of self-report color-blindness and failure to report seeing numbers on the critical trials.

2.1. Method

2.1.1. Subjects

We recruited 310 workers on TurkPrime, with a HIT posting entitled "Short survey on visual perception. Must be color-blind". The HIT was only available to workers who: (a) reside in the United States, (b) have completed at least 100 MTurk tasks, and (c) have a minimum approval rating of 50%. Our sample size is comparable to other published studies that aimed to evaluate the methodological aspects of MTurk (e.g., Buhrmester et al., 2011; Chandler et al., 2014; Hauser & Schwarz, 2016; Siegel, Navarro, & Thomson, 2015).

2.1.2. Material

A 4-item demographic questionnaire was used to assess age, gender, color-blindness status, and type of color-blindness. The age question was open-ended (i.e., "How old are you?") and the other questions were forced choice. Wording of the questions, response options, and a summary of the responses are presented in Table 1. Two of the demographic questions are central to our investigation. First, responses to the "Are you color-blind?" question afford the opportunity to identify those individuals who ignored the explicit eligibility requirement and those who self-reported to be color-blind. Second, answers to the type of color-blindness question served as an additional index of likely deception. We purposefully included a fictitious type of color-blindness (i.e., Red/Blue) as a possible option; we contend that only people with little knowledge about color-blindness would select that option and endorsement would reflect purposeful demographic deception (or inattentive responding as a "best case" scenario).

A total of 10 color plates were included. Of greatest importance are the two critical plates, where numbers are visible to *all* individuals, regardless of color-vision integrity. One of these critical stimuli was selected from the Ishihara test (Ishihara, 1972), and the other critical stimulus was selected from the Dvorine Pseudo-Isochromatic Plates (Dvorine, 1953), another common color vision test. We further selected eight test plates from the Ishihara test to be used as filler trials; each plate contains a number that is visible only to individuals with normal color vision. Since the sole purpose of these filler plates is to increase the length of the task, we will not consider those responses in our analyses.

Table 1
Demographic questions, response options and summary of responses in Study 1. Total number of respondents: 310.

Demographic question & response options	Percentage (and number) of respondents
How old are you?[a]	
18 - 25	23.2% ($n = 72$)
26 - 30	30.6% ($n = 95$)
31 - 35	19.4% ($n = 60$)
36 - 40	11.0% ($n = 34$)
41 - 50	10.0% ($n = 31$)
51 - 60	3.5% ($n = 11$)
62 - 67	1.9% ($n = 6$)
75	0.3% ($n = 1$)
What is your gender?	
Female	36.8% ($n = 114$)
Male	62.6% ($n = 194$)
Do Not Wish to Say	0.6% ($n = 2$)
Other	0.0% ($n = 0$)
Are you color-blind?	
Yes	69.4% ($n = 215$)
No	30.6% ($n = 95$)
What form of color-blindness do you have?	
Red/Blue[b]	15.8% ($n = 49$)
Red/Green	28.1% ($n = 87$)
Yellow/Blue	6.8% ($n = 21$)
Total color-blindness	10.7% ($n = 33$)
Not sure	8.1% ($n = 25$)
Not color-blind	30.6% ($n = 95$)

[a] The age question is open-ended, but responses are grouped here for summary purposes only.
[b] "Red/Blue" color-blindness is fictitious and is included as an index of worker deception.

2.1.3. Procedure

The eligibility requirement of "must be color-blind" was explicitly stated in both the HIT title and in the consent form. Once workers accepted the HIT, they were redirected to the survey hosted on Qualtrics. After completing an online consent form, workers answered the four demographic questions described in Table 1. Before beginning the visuo-perceptual task, workers were asked to disable any assistive technology that they may use on their computers to accommodate their color-blindness.[2]

For the visuo-perceptual task, workers were informed that they would encounter a series of images, with each image consisting of colored dots. In some cases, a number (e.g., "14") is visible among the colored dots, and in other cases, there is no number. If they saw a number, even if they were not 100% sure of its identity, they should enter the number in the response box. And if they did not see a number, they should select the "I do not see a number, just dots" option. The two critical trials were presented on trials 1 and 6.

At the end of the session, workers received a completion code that they would provide to MTurk as proof of participation. Each session lasted approximately 3 min, and workers who provided the correct completion code received $0.25 in their MTurk accounts. All procedures were approved by Villanova University's Institutional Review Board.

2.2. Results

2.2.1. Compliance to stated eligibility requirement

As presented in Table 1, 30.6% (95 out of 310) of our sample ignored the eligibility requirement (i.e., "must be color-blind") that was explicitly stated in both the HIT title and the consent form. Despite the fact that these rule breakers ignored the inclusion criterion, they were consistent in their responding, in that all 95 workers also selected "Not color-blind" in the type of color-blindness question. Furthermore, all workers in this subgroup responded honestly on the critical trials, such that all workers reported the correct numbers on both critical trials. In sum, we conclude that these workers simply ignored the stated eligibility requirement and did not attempt to lie about their demographic.

2.2.2. Demographic misrepresentation

As described earlier, the primary index of demographic misrepresentation is the responses on the critical trials. To reiterate, the critical trials consist of color plates where *all* individuals, regardless of color vision integrity, should be able to see the embedded numbers within the colored dots. We contend that imposters would attempt to maintain the façade of being color-blind by relying on their faulty beliefs about the condition when responding. Thus, we categorize a respondent as an imposter if he or she self-identifies as color-blind and fails to correctly identify the numbers on the critical trials. Of the 215 individuals who claimed to be color-blind, 42.8% (n = 92) failed to report the correct numbers on the critical trials. This finding is consistent with Wessling, Huber

[2] We do not expect subjects' compliance to this request to impact the interpretation of our findings. Again, the critical trials consist of numbers that are visible to all individuals, regardless of color vision integrity. Thus, the absence or presence of color assistive technology should not actually impact performance.

and Netzer's (2017) observation that imposters tended to rely on stereotypes when responding to questions to which they have little personal knowledge.

We also included a fictitious type of color-blindness as a supplementary measure of demographic deception. Of those who claimed to be color-blind, 22.8% (n = 49) reported to experience Red/Blue color-blindness, a condition that does not exist. The endorsement of the fictitious Red/Blue variant is even more remarkable considering workers had the option to report "Not sure" when inquired about the type of color-blindness they experience.

Thus, the total percentage of imposters in our sample is likely to be 55.8%, if we include those who failed to report the correct numbers on both critical trials (n = 92) and those who reported correct numbers on both critical trials but claimed to have a condition that does not exist (n = 28).

2.2.3. Characteristics of rule breakers, honest respondents, and imposters

A one-way analysis of variance revealed a significant age difference across the three groups of respondents: honest respondents ($M = 34.3$, $SD = 11.1$), imposters ($M = 29.6$, $SD = 8.3$), and rule breakers ($M = 33.6$, $SD = 9.6$). Tukey HSD tests revealed that imposters are significantly younger than both honest respondents and rule breakers, whereas the honest respondents and rule breakers did not differ in age.

2.3. Discussion of downstream consequences of worker deception

A few noteworthy downstream consequences can be gleaned from our data. If we had relied on our sample to inform our knowledge about color-blindness, we would have drawn the following conclusions. First, based on our sample, the relative prevalence of color-blindness in men and women was approximately 2.2 to 1 across all sub-types and roughly 3.6 to 1 in red/green color-blindness. This is drastically different from the estimated male to female ratio of 16 to 1 in red/green color-blindness reported by the National Eye Institute (NEI) in 2015. Second, our sample revealed a never before reported form of red/blue color-blindness, and it was the second most common form of color-blindness in our sample. Third, we also would have concluded the following prevalence ordering of color-blindness types, with the most common being red/green, followed by red/blue, then complete color-blindness, yellow/blue and unsure, a rank order that differed from that reported by the NEI (even if we eliminated the fictitious red/blue type). According to the NEI, the three most common forms of color-blindness, in order of prevalence, are: red/green, yellow/blue, and complete color-blindness. In sum, any conclusions based on these findings would have jeopardized the scientific integrity of the investigation and led to misleading generalizations concerning this genetic condition.

3. Study 2

Given the paucity of evidence that demonstrates the breadth of downstream consequences of worker deception, coupled with the fact that these consequences could pose a serious threat to scientific conclusions, we sought to investigate other ways in which these consequences may manifest themselves. Since these effects have yet to be investigated in the context of more complex psychological constructs, we sought to do so in Study 2. Specifically, we utilized the Future Time Perspective Scale (FTP Scale, Carstensen & Lang, 1996), a 10-item survey designed to measure an individual's perceived time horizon, that is, their subjective sense of time until death (e.g., Carstensen, 2006; Fung & Isaacowitz, 2016). We chose this scale because its well-established psychometric and psychological properties would allow us to examine the effects of demographic deception on data quality in two ways. First, the factor structure of the FTP Scale has been established in previous studies (e.g., Cate & John, 2007; Kozik, Hoppmann, & Gerstorf, 2015), and we can leverage this knowledge to assess whether imposters would be more likely to provide haphazard responses than honest respondents. By comparing the latent structure in our samples' responses, we can assess whether honest respondents and imposters are similarly attentive to the task demands. Second, since it is well-established in the aging literature that as one ages, time perspective becomes more limited (for reviews, see Carstensen, Fung, & Charles, 2003; Fung & Isaacowitz, 2016), we investigated whether this pattern would persist in imposters. If so, it would suggest that imposters' deception is limited to participation eligibility. If not, it would suggest that imposters' responses are qualitatively different from those provided by honest respondents. While both of these alternatives pose a threat to data quality, they represent distinct risks to scientific integrity. In light of the rising utilization of online data collection platforms by both survey and experimental researchers in Psychology (e.g., Crump et al., 2013; Heer & Bostock, 2010; Paolacci et al., 2010; Stewart et al., 2017), an examination of this sort is crucial.

In this study, we compared workers' self-report demographic characteristics across two sessions. For session 1, we recruited workers by posting a series of tasks with specific inclusion criteria (e.g., females only), and we included a total of five demographic variables that are commonly used to determine eligibility in psychological research. Several days later, we invited all session 1 workers to complete a different task (session 2) that did not have any eligibility requirement. We reasoned that if workers were motivated to maximize task completions and enroll in as many studies as possible, including those for which they did not qualify, then session 1 should elicit purposeful misrepresentation in these individuals. However, when there was no threat to participation eligibility, as in session 2, workers would have little incentive to misrepresent themselves and would likely provide truthful demographic characteristics. Thus, discordant demographic responses between the two sessions can be taken as an index of misrepresentation.

3.1. Method

3.1.1. Subjects

We recruited 502 workers for session 1. Our HITs were only visible to those workers who (a) did not participate in Study 1, (b) reside in the United States, (c) had completed at least 100 MTurk tasks, and (d) had a minimum approval rating of 50%. All workers responded to the HITs that we posted on TurkPrime's website. The HITs were posted consecutively to ensure that workers could not participate in more than one HIT.

Approximately 48–72 h later, we used TurkPrime's email system to send a message to all session 1 respondents, inviting them to participate in a second, unrelated study (session 2). A total of 463 of the original workers took part in session 2. As described earlier, since session 1 was designed to elicit misrepresentation of demographic characteristics and that session 2 responses were more likely to be truthful, we summarized our sample's characteristics based only on their session 2 responses (see Table 2).

3.1.2. Design

We examined five demographic variables in this study: age, gender, education, income, and family status. We chose these attributes because they represent commonly used selection criteria for theoretically driven research (e.g., targeting older adults for

Table 2
Summary of demographic responses from session 2: Condition (e.g., age, gender, education, income, family status), number of workers who participated in both sessions for each condition, demographic questions, response options, and percent of subjects who selected the respective responses. Total number of workers who participated in both sessions: 463.

Condition *(n)*	Demographic question & response options	Percentage of workers who selected each option within each condition
Age ($n = 95$)	What year were you born?	
	1939 or earlier	0.0%
	1940–1949	0.9%
	1950–1959	7.1%
	1960–1969	11.4%
	1970–1979	23.3%
	1980–1989	37.8%
	1990–1998	19.4%
Gender ($n = 91$)	What is your gender?	
	Female	53.8%
	Male	45.8%
	Do not wish to say	0.4%
Education ($n = 93$)	What is your level of education?	
	Did not complete high school	0.4%
	High school graduate/GED	30.5%
	Associate degree	21.2%
	College graduate	36.1%
	Post-college education	11.9%
Income ($n = 95$)	What is your annual household income?	
	Less than $14,999	10.2%
	$15,000 to $24,999	13.2%
	$25,000 to $39,999	16.6%
	$40,000 to $59,999	24.8%
	$60,000 to $74,999	14.0%
	$75,000 to $99,999	10.6%
	$100,000 or above	10.6%
Family status ($n = 89$)	What is your family status?	
	Single	45.1%
	Married/living with a partner	54.9%

aging studies). Furthermore, some of these characteristics were not investigated in the studies reviewed earlier, thus allowing us to replicate and extend previous findings.

Within each condition, we included two levels, with each level representing a different inclusion criterion. As such, a total of 10 different between-subject HITs were created. Each session 1 HIT was advertised with the identical title, and the eligibility requirement for each condition was also specified (e.g., “Outlook on Life, Females only”; “Outlook on Life, must be born in the 1970s”). The eligibility requirement was repeated in the consent form. The 10 HITs were posted on the website one at a time, and exclusion criteria were set such that each worker could participate in only one assignment.

Approximately 48–72 h after session 1, we sent an email to invite all session 1 subjects to participate in an unrelated study entitled “Aesthetic Preferences”. The study was only open to individuals who already participated in session 1, and importantly, in contrast to session 1, this study did not have any stated eligibility requirements. The email included the researcher's identity, the survey's link and a subject line that comprised the title, duration and pay. Two additional reminders, spaced approximately 24 h apart, were sent to subjects who did not respond to a previous invitation. Although the workers knew the identity of the researcher, they were not informed of the relationship between the two studies, as the email simply stated that a new study was available to them. In other words, from the workers' perspective, the two sessions were unrelated. Since the two studies were perceived as unrelated and were spaced several days apart, we believe it would be improbable for workers to remember the eligibility specifications from the first session. Nonetheless, in the unlikely event that they did remember and tried to provide the same demographic responses as session 1, the estimates we provided below would actually reflect an underestimation of misrepresentation.

3.1.3. *Material*

Workers responded to the same demographic questions in both sessions (see Table 2). In session 1, we also administered the FTP Scale (see Table 3 for items; Carstensen & Lang, 1996), a 10-item survey designed to measure an individual's perceived time horizon (e.g., Carstensen, 2006; Fung & Isaacowitz, 2016). As described earlier, the primary goal of session 2 was to ascertain truthful demographic information. Nonetheless, we needed a cover task to mask the purpose of session 2 and its relation to session 1. We opted for an aesthetic judgment task that required workers to indicate their preference for 10 Impressionist paintings that we gathered from the Internet. Since this task employed procedures that are typical of MTurk studies and was distinct from the FTP scale, we reasoned that it should help minimize any suspicions respondents may have about the relationship between the two sessions.

3.1.4. *Procedure*

Once workers accepted the HIT, they were redirected to the survey hosted on Qualtrics. After completing an online consent form, each worker provided his/her MTurk worker ID, an alphanumeric code that is unique to each individual. Collection of the IDs allowed us to collate individual subject's data across the two sessions. The workers then answered the five demographic questions listed in Table 2, which were identical across the two sessions.

In session 1, statements from the FTP scale were presented immediately after the demographic questions, and subjects rated how well each statement described him/her using a Likert scale that ranged from 1 (very untrue for me) to 7 (very true for me). In session 2, the Impressionist paintings were presented immediately after the demographic questions, and subjects rated how much they liked each painting using a 5-point Likert scale.

At the end of each session, workers received a completion code that they would provide to MTurk as proof of participation. Each session lasted approximately 3 min, and workers who provided the correct completion code received $0.25 in their MTurk accounts. All procedures were approved by Villanova University's Institutional Review Board.

3.2. *Results*

We restricted our analyses to those workers who completed both sessions (463 out of 502 workers), as we will not be able to ascertain the veracity of their demographic responses if they only participated in session 1.

3.2.1. *Compliance to stated eligibility requirement*

We defined non-compliance as selection of responses that were incompatible with the stated eligibility requirements in session 1 (e.g., selected “born between 1950 and 1959” in the “born in the 1970s” condition). Overall, 13.6% (63 out of 463) of our sample fell into this category, where workers made no attempt to provide misleading information in order to fulfill the stated eligibility criterion. We interpret these responses as a sign of either blatant disregard of or inattention to the eligibility requirements.

Table 3
Factor loading comparison for the FTP Scale. For the sake of simplicity, we reported the name of the factor with the highest loadings for prior work (see Table 4, Study 2 for Cate & John, 2007 and Table 2 for Kozik et al., 2015 for factor loading values). Factor loading values from the current study are listed below, with "Opp" referring to "Opportunities" and "Lim" referring to "Limitations". Finally, the highest loading for each item is shown in bold.

Item	Statement	Cate and John (2007)	Kozik et al. (2015)	Current study: Imposters		Current study: Honest respondents	
				Opp	Lim	Opp	Lim
1	Many opportunities await in the future	Opportunities	Opportunities	**.91**	.10	**.89**	.21
2	I expect that I will set many new goals in the future	Opportunities	Opportunities	**.87**	.18	**.89**	.13
3	My future is filled with possibilities	Opportunities	Opportunities	**.93**	.09	**.90**	.20
4	Most of my life lies ahead of me	Opportunities	Opportunities	**.68**	.38	**.72**	.43
5	My future seems infinite to me	Opportunities	Opportunities	**.76**	.37	**.61**	.49
6	I could do anything I want in the future	Opportunities	Opportunities	**.85**	.23	**.70**	.39
7	There is plenty of time left in my life to make new plans	Opportunities	Opportunities	**.74**	.47	**.69**	.49
8	I have the sense that time is running out	Limitations	Limitations	.06	**.85**	.26	**.85**
9	There are only limited possibilities in my future	Limitations	Limitations	.58	**.62**	.50	**.64**
10	As I get older, I begin to experience that time is limited	Limitations	Limitations	.22	**.80**	.12	**.88**

3.2.2. *Demographic misrepresentation*

We excluded data from workers who disregarded the explicit eligibility requirement, and this resulted in a final group of 400 respondents. We quantified misrepresentation by contrasting the responses to the critical demographic question across the two sessions. For example, we compared the age responses in the two sessions for subjects in the age conditions and compared the income responses in the two sessions for workers in the income conditions. Since session 2 did not have any stated eligibility requirements, we reasoned that workers would have little incentive to be deceitful and that their responses are most likely to reflect their true demographic characteristics. Thus, any discrepancy in their responses to the critical demographic question between the two sessions can be interpreted as an attempt to misrepresent themselves in order to qualify for a study. Overall, 21.8% of our sample provided mismatched responses.

3.2.3. *Variability in deception rates*

Another striking aspect of our results is the variability in the rate of misrepresentation across the demographic conditions. Specifically, the following percentage of respondents in each condition provided discrepant responses to the critical demographic question between the two sessions: Age = 22.6%, Education = 31.3%, Gender = 6.6%, Income = 38.2%, Family Status = 14.8%. One potential source for this variability may be the relative proportion of individuals who would be deemed ineligible if they were to abide by the inclusion criterion. Preliminary support for this possibility comes from published demographic norms on http://demographics.mturk-tracker.com. According to data on this website, the family status and income criteria we used excluded ~42% and ~77% of the MTurk population, respectively. Our data revealed ~15% and ~38% misrepresentation in those two conditions, respectively. Based on our sample, we speculate that as the proportion of ineligible workers increases, the proportion of individuals who are likely to misrepresent themselves also increases. This notion is also consistent with Wessling, Huber and Netzer's (2017) finding that the more stringent the inclusion criterion or the more rare the target sample (e.g., owning a kayak), the greater the percentage of deception (see also Chandler & Paolacci, 2017).

3.2.4. *Downstream consequences: comparisons of response patterns between imposters and honest respondents*

We examined the downstream consequences of worker deception in two ways. In the first set of analyses, we asked whether imposters are more likely to provide haphazard responses than honest respondents. We capitalized on prior results that have established the factor structure of the FTP scale (e.g., Cate & John, 2007; Kozik et al., 2015), with the primary factors being "Opportunities" (items 1 through 7) and "Limitations" (items 8 through 10). We hypothesized that if imposters in our study were inattentive during the task, their responses to the FTP Scale would also be random, which would be confirmed by a factor structure that differed from that reported in previous studies and that observed in the honest respondents in our sample. Alternatively, if the imposters' deceitful behavior is limited to gaining entry into the study and they actually responded to the task attentively, then we would expect the patterns of their responses to reveal similar factor loadings as those found in prior work and in the honest respondents in our sample.

First, we explored whether the factor loadings from our samples would echo those reported by Cate and John (2007) and Kozik et al. (2015). Table 3 summarized the findings from our exploratory factor analyses. We conducted two Principal Component Analyses, one for each group and found that both groups revealed the same overall pattern: Factor 1 ("Opportunities") explained 61.8% of variance (eigenvalue = 6.18) in the imposter group and 62.4% of variance (eigenvalue = 6.24) in the honest respondent group, and Factor 2 ("Limitations") explained 13.0% of variance (eigenvalue = 1.30) in the imposter group and 11.9% of variance (eigenvalue = 1.19) in the honest respondent group. A visual inspection of these patterns suggest that both groups in our sample responded in a way that revealed the same factor loadings as prior work (Cate & John, 2007; Kozik et al., 2015) and that both groups appeared to have responded to the FTP scale in a similarly meaningful way.

To confirm that responses from the honest respondents and the imposters indeed revealed similar latent structures, we conducted a multi-group exploratory factor analysis, an outlined in van de Schoot, Lugtig and Hox (2012). As suggested by van de Schoot, Lugtig and Hox and also Burnham and Anderson (2004), we selected the most stringent invariance model (where factor loadings and intercepts are invariant) because it yielded the lowest Bayesian Information Criterion value (see Table 4 for model comparisons and fit indices). In other words, the multi-group exploratory factor analysis confirmed similar latent structures between the honest respondents and the imposters.

Although the finding of measurement invariance suggests that imposters are just as attentive to the task as honest respondents, it does not address whether their responses are psychologically meaningful. In other words, even though we know that both groups were responsive to the task demands (i.e., responded in such a way as to reveal that the scale assesses "Opportunities" and "Limitations"), it remains unclear how carefully they considered each statement in relation to their own lives (i.e., how does *my* future look at this point in my life, are there many opportunities awaiting

Table 4
Summary of model comparison for multi-group exploratory factor analysis. Model with lowest BIC is highlighted in bold font.

Model[a]	ChiSq	df	p	CFI	TLI	RMSEA	BIC	AIC
1	182	52	<.0001	0.958	0.928	0.112	12,795	12,484
2	209	68	<.0001	0.955	0.940	0.102	12,726	12,479
3	218	62	<.0001	0.950	0.928	0.112	12,771	12,499
4	**238**	**76**	**<.0001**	**0.948**	**0.938**	**0.103**	**12,707**	**12,492**

[a] Model 1: No constraints. Model 2: Factor loadings invariant. Model 3: Intercepts invariant. Model 4: Factor loadings and intercepts invariant.

me?) To explore this issue, we examined the relationship between age and the FTP score. As reviewed above, the relationship between age and future time perspective is well-established, such that as one ages, their time perspective also becomes more constrained (for reviews, see Carstensen et al., 2003; Fung & Isaacowitz, 2016). We examined this relationship in the imposters and in the honest respondents.

We followed published scoring procedure (Stanford Life-span Development Laboratory, 2016) and derived an FTP score for each subject, with higher scores indicating perceived expansiveness of time horizon (Carstensen & Lang, 1996). Thus, a negative relationship between age and FTP score is expected. Scores from the two groups of respondents spanned a similar range (Imposters: 1.5 to 7.0; Honest respondents: 1.1 to 7.0), and, the mean scores for the two groups were also similar (Imposter $M = 4.4$, $SD = 1.4$; Honest respondents $M = 4.6$, $SD = 1.3$, $t\,[398] = 0.136$, $p = .892$).

We conducted a moderator analysis to determine whether the relationship between age and FTP score differed by deception status. Specifically, we conducted a multiple regression analysis with “FTP score” as the dependent variable, and “age”, “deception status”, and the interaction term “age x deception status” as predictors. Since we do not have respondents' exact ages, we used the decade in which they were born as a proxy for age. Thus, age was coded as an ordinal variable, and deception status as a categorical variable. Replicating established results, we found that age significantly predicted FTP score ($\beta = -.269$, $p < .001$). However, this relationship was not moderated by deception status, as confirmed by the minimal R^2 change (0.1%) in the model that included the “age x deception status” interaction term as a predictor. Taken together, these findings suggest that deception status did not have a substantive impact on the relationship of interest.

3.3. Discussion

Two key findings emerged from Study 2. (a) We found that close to 22% of our sample provided discrepant demographic information across the two sessions, and we interpreted this discrepancy as an indication of purposeful misrepresentation. (b) We found that responses from both imposters and honest workers revealed a similar latent structure in a measure of a complex psychological construct. Furthermore, we replicated a well-established finding in the aging literature (see Carstensen et al., 2003 for a review), where both groups revealed an age-related shift in time perspective. Implications of these findings will be considered in the General Discussion.

4. General discussion

As behavioral researchers become increasingly reliant on MTurk for data collection, it is critical that we identify its benefits and shortcomings. Across two studies, we focused on demographic misrepresentation and the potential downstream consequences that may result from such deception. Although the notion that online interactions encourage deception is not new (see Ott, Cardie, & Hancock, 2012; Toma, Hancock, & Ellison, 2008 for examples in the context of online dating and online reviews, respectively), it has not been evaluated systematically in the context of MTurk workers until very recently (Chandler & Paolacci, 2017; Wessling et al., 2017). To contribute to this growing literature, we asked, “Are we testing who we think we are testing?” and “Does demographic deception ultimately have an impact on data quality?”

4.1. Demographic misrepresentation: are we testing who we think we are testing?

Under conditions of explicit eligibility requirements, an alarmingly high proportion of respondents who chose to complete our HITs provided misleading demographic information in order to qualify for the studies (55.8% in Study 1 and 21.8% in Study 2). These disturbingly high misrepresentation rates raise concerns about the generalizability and validity of MTurk data. This pattern poses a particular challenge to those researchers who rely exclusively on MTurk data to draw conclusions about the behaviors of specific demographic groups. Consider the situations of a consumer researcher interested in how men would respond to an advertisement campaign for a beard trimmer or a psychologist interested in the symptoms of post-partum depression. By definition, the researchers would want to restrict their studies to those specific demographic groups (i.e., men with beard and women who recently experienced childbirth, respectively), but based on our findings, selection of respondents based on self-report eligibility could result in a considerable proportion of the data coming from individuals who do not fit the desired selection criterion. Faulty generalizations based on these inappropriate data would seriously threaten scientific integrity.

One could argue that the high deception rate and downstream consequences we observed may be due to the fact that we have a liberal criterion of approval ratings for worker selection, allowing those with an approval rating of 50% or higher to see the HITs. In other words, maybe we are allowing all the bad apples (those workers with low approval ratings) into the study and those workers are biasing the results. This explanation seems unlikely for two reasons. First, Chandler and Paolacci (2017) established that worker quality has little influence over deception rate (Study 4a). Second, according to an Amazon representative (S. Krumholtz, TurkPrime Account Manager, personal communication, July 31, 2017), most workers possess approval ratings of 90% or higher. As such, even with a 50% approval rating criterion, most workers who participated in our tasks most likely have approval ratings of 90% or higher. Thus, it seems unlikely that our results can be fully explained by worker quality.

One potential limitation of our studies is that we do not have in-lab comparison groups, so it remains unclear whether the high deception rates are unique to the online labor force. We speculate that these rates would be lower for laboratory studies because the efforts required on the part of the participants are higher and the likelihood of being caught lying is also higher.

Given previous findings that workers are generally honest unless incentivized to lie (e.g., Chandler & Paolacci, 2017; Chandler & Shapiro, 2016; Rand, 2012; Wessling et al., 2017), the apparent solution seems to be an elimination of such eligibility requirements in online studies. However, this seemingly obvious solution places other impractical constraints on researchers who have theoretically motivated reasons to target a specific population. Indeed, it would be a waste of resources to conduct a study without restrictions if the researchers' inquiry can only be addressed by a specific demographic group. Finally, our data also suggest that a simple reliance on respondents to comply with eligibility requirement is insufficient, as 30.3% of our sample in Study 1 and 13.6% of our sample in Study 2 failed to comply.

4.2. Downstream consequences: does demographic deception ultimately have an impact on data quality?

Another important contribution of our work is an evaluation of potential downstream consequences of demographic deception. In other words, if the imposters were willing to lie in order to qualify for a study, would their deceitful behaviors have ramifications on the rest of the study, thus threatening data quality? We found differing effects of demographic deception on data quality in our studies.

In Study 1, we evaluated the impact of demographic deception in the context of a genetic condition. We found that almost 56% of respondents to our HIT were imposters. Consistent with Wessling et al. (2017), we also found that imposters tended to rely on generic knowledge when they lacked personal experience with the condition. In this case, they relied on a common misconception about color-blindness, which revealed itself in the critical trials of the visuo-perceptual task. As discussed in section 2.3, the conclusions we would have drawn about color-blindness based on those data would have been erroneous and misleading, and such faulty generalizations would have seriously jeopardized the scientific integrity of the investigation.

In Study 2, we asked whether deception status would influence worker attentiveness and the psychological validity of their responses. Surprisingly, we found very similar patterns of results across the two groups, where responses from both imposters and honest respondents revealed similar psychometric and psychological properties of the FTP Scale. These findings suggest that deception status may not impact data quality in this context.

Taken together, the nature of downstream consequences of demographic deception differed between our two studies. Indeed, the effects of worker deception seem far more damaging and concerning in Study 1 than in Study 2. What are some factors that may contribute to this divergence?

One possibility is that the inherent interest and engagement level differ between the two tasks. We speculate that workers are more likely to find contemplating future time perspective more interesting and engaging than identifying numbers among colored dots, as the former task is more self-relevant. Consequently, it is plausible that workers are more likely to respond genuinely when they are more engaged with the material. This notion is consistent with the observation that many workers reported enjoyment as a key motivating factor for taking part in MTurk studies (Buhrmester et al., 2011). Relatedly, it remains an open question whether imposters' responses would differ from honest respondents if the tasks were longer, more effortful, or less engaging.

A more plausible explanation may rest with the relationship between the demographic variable of interest and the outcome measure. For instance, in Study 1, the critical trials in the visuo-perceptual task target the workers' personal experience with colorblindness. That is, workers who pretended to be colorblind would rely on their misconception about the condition and provide incorrect responses, thus revealing their deception. However, in Study 2 we evaluated the extent of demographic deception across a range of demographic variables that are commonly used to determine eligibility in psychological studies. While this information extends our understanding of the nature of demographic misrepresentation, it might have obscured the effects of deception on an age-related complex psychological measure. In other words, given that the relationship of interest is that between age and future time perspective, it is conceivable that only those imposters who lied about their age would impact data quality. Thus, a more instructive analysis approach may be to compare patterns of responding between imposters in the age condition (those who misrepresented their age in order to qualify for the study) and imposters in the other conditions (those who misrepresented themselves in a non-age demographic category). However, the relatively small number of imposters in the age condition ($n = 19$) renders such an analysis under-powered and inappropriate.[3] Future studies could address this limitation by improving the match of the relationship between the demographic variable of interest and the outcome measure used to identify imposters.

In sum, our findings are consistent with the notion that explicit prescreening criteria could have significant ramifications on data quality (Chandler & Paolacci, 2017; Kees et al., 2017; Smith, Roster, Golden, & Albaum, 2016; Wessling et al., 2017). Through the use of two distinct tasks, we have further specified the nature of these downstream consequences.

4.3. Balancing the benefits and drawbacks of online data collection

The obvious question thus becomes: How do we balance MTurk's advantages with these significant flaws that threaten generalizability and data quality? Many researchers have provided excellent comprehensive tutorials with strategies to improve data quality (for a few recent examples, see Chandler & Paolacci, 2017; Chandler & Shapiro, 2016; Goodman & Paolacci, 2017; Kees et al., 2017; Stewart et al., 2017; Wessling et al., 2017). Specific to the issue of worker deception for the purpose of study eligibility, the consensus appears to be that the best way to minimize deception is to avoid explicit prescreening because workers are motivated to lie only when incentivized, such as when their ability to participate in studies is threatened. However, as discussed earlier, there are often theoretically-motivated reasons to target a specific demographic group, such as in cases of clinical research (Chandler & Shapiro, 2016). The alternative of administering a study without requirements and then excluding the data from ineligible workers post-hoc seems ineffective, and it will likely negate the cost-effectiveness advantage if the sample of interest is rare. Here, we offer two related approaches that may ameliorate the issue of worker deception.

One possibility is to utilize Amazon's Panel Study[4] to restrict workers' access to studies based on specific demographic criteria (since these data are provided at the time of account setup, it is reasonable to assume that the responses would be truthful). However, the fee for this premium service may diminish MTurk's cost-effectiveness advantage. For example, including gender and age criteria would more than double the per-worker cost. Furthermore, if the selection criterion is more specific (e.g., having a specific medical condition), that information is unlikely to be available via this service.

A similar approach is for researchers to offer micropayment for completion of a customized pre-study demographic questionnaire

[3] Nonetheless, there are hints that the findings may indeed differ between age imposters and other demographic imposters. Whereas age imposters failed to show a significant relationship between age and FTP score, r_s $(17) = -.152$, $p = 0.535$, other imposters revealed a marginally significant negative correlation between age and FTP score, $r\ (66) = -0.225$, $p = .065$. Again, given the small number of subjects, these suggestive patterns must be interpreted with caution.

[4] It should be noted that Amazon's Panel Study is different from professional panels, which are coordinated by third-party companies that typically receive financial incentives for identifying participants that meet certain selection criterion. Thus, researchers are likely to have less control over the participant selection process used by these third-party companies. Indeed, some researchers have questioned data quality from participants in professional panels (e.g., Downes-Le Guin, Mechling, & Baker, 2006; Kees et al., 2017).

and later select subjects based on those responses (Paolacci et al., 2010; Springer, Martini, Lindsey, & Vezich, 2016; Wessling et al., 2017). Although this alternative is more cost-effective, it is potentially time-consuming. Given that the composition of the MTurk workforce is constantly evolving, researchers may need to re-administer the pre-study demographic survey regularly to capture the most recent cohort of available workers. Similarly, researchers interested in characteristics that are less stable, such as product ownership, will also need to re-administer the pre-study survey frequently (Chandler & Paolacci, 2017; Wessling et al., 2017).

Future research would benefit from investigations that aimed to understand the factors that influence demographic deception. Aside from monetary gain, what are other factors that may motivate workers to lie? How does characteristic stability affect rate of deception? Is it more likely for individuals to be deceptive about characteristics that they could imagine (e.g., A dog owner pretending to own a cat)? How will an imposter's desire to maintain the appearance of authenticity bias the results? In what way does the duration and engagement level of the task impact the downstream consequences of demographic deception?

In sum, we propose that the advantages of MTurk may come with a cost that could threaten generalizability and validity, and the promise of half a million workers from 190 countries is empty if we cannot be sure whom we are really testing. Our findings also suggest that when considering data quality from the MTurk workforce, we must look beyond whether the patterns of results are comparable between MTurk and in-lab samples and must also consider other metrics, such as demographic validation. Although these concerns may be minimized in studies that do not make claims about specific demographic groups, we urge researchers to be mindful of these flaws and be judicious in their designs, claims and conclusions.

Acknowledgments

We thank Dr. Rebecca Brand for helpful comments on an earlier draft of the manuscript and Dr. Janette Herbers and Dr. John Kurtz for guidance on data analyses. This research was supported by a research grant from the Center for Marketing and Consumer Insights at Villanova University School of Business.

References

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602.

Bates, J. A., & Lanza, B. A. (2013). Conducting Psychology student research via the Mechanical Turk crowdsourcing service. *North American Journal of Psychology, 15*(2), 385–394.

Bohannon, J. (2011). Social science for pennies. *Science, 334*(6054), 307. https://doi.org/10.1126/science.334.6054.307.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. https://doi.org/10.1177/1745691610393980.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304. https://doi.org/10.1177/0049124104268644.

Carstensen, L. L. (2006). The influence of a sense of time on human development. *Science, 312*, 1913–1915. https://doi.org/10.1126/science.1127488.

Carstensen, L. L., Fung, H. H., & Charles, S. T. (2003). Socioemotional selectivity theory and the regulation of emotion in the second half of life. *Motivation and Emotion, 27*, 103–123. https://doi.org/10.1023/A:1024569803230.

Carstensen, L. L., & Lang, F. R. (1996). *Future time perspective scale*. Retrieved from http://psych.stanford.edu/~lifespan/links.htm.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009.

Cate, R. A., & John, O. P. (2007). Testing models of the structure and development of future time perspective: Maintaining a focus on opportunities in middle age. *Psychology and Aging, 22*, 186–201. https://doi.org/10.1037/0882-7974.22.1.186.

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

Chandler, J., & Paolacci, G. (2017). *Lie for a Dime: When most prescreening responses are honest but most "eligible" respondents*. Social Psychological and Personality Science.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science, 26*, 1131–1139. https://doi.org/10.1177/0956797615585115.

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

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410.

Deneme. (2016). *Deneme: A blog of experiments on Amazon Mechanical Turk*. Retrieved September 20, 2016, 2016, from http://groups.csail.mit.edu/uid/deneme/.

Downes-Le Guin, T., Mechling, J., & Baker, R. (2006). Great results from ambiguous sources: Cleaning internet panel data. In *ESOMAR world research Conference: Panel research*.

Dvorine, I. (1953). *Dvorine pseudo-isochromatic plates*. New York: The Psychological Corporation.

Fung, H. H., & Isaacowitz, D. M. (2016). The role of time and time perspective in age-related Processes: Introduction to the special issue. *Psychology and Aging, 31*(6), 553–557. https://doi.org/10.1037/pag0000119.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing of consumer research. *Journal of Consumer Research, 44*, 196–210. https://doi.org/10.1093/jcr/ucx047.

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

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

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 100–407. https://doi.org/10.3758/s13428-015-0578-z.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Paper presented at the SIGCHI conference on human factors in computing systems, Atlanta, Georgia*.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature, 466*, 29. https://doi.org/10.1038/466029a.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83.

Highhouse, S., & Zhang, D. (2015). The new fruit fly for applied psychological research. *Industrial and Organizational Psychology, 8*(2), 179–183. https://doi.org/10.1017/iop.2015.22.

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

Ishihara, S. (1972). *Tests for colour-blindness* (24 Plates Edition ed.). Tokyo, Japan: Kanehara Shuppan Co., Ltd.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141–155. https://doi.org/10.1080/00913367.2016.1269304.

Kozik, P., Hoppmann, C. A., & Gerstorf, D. (2015). Future time perspective: Opportunities and limitations are differentially associated with subjective well-being and hair cortisol concentration. *Gerontology, 61*, 166–174. https://doi.org/10.1037/0882-7974.22.1.186.

Lease, M., Hullman, J., Bigham, J. P., Bernstein, M. S., Kim, J., Lasecki, W., … Miller, R. C. (2013). *Mechanical Turk is not anonymous*. Social Science Research Network (SSRN). https://doi.org/10.2139/ssrn.2228728. https://ssrn.com/abstract=2228728.

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for behavioral sciences. *Behavior Research Methods*. https://doi.org/10.3758/s13428-016-0727-z. online.

National Eye Institute, NEI. (2015). Retrieved August 10, 2017, from https://nei.nih.gov/health/color_blindness/facts_about.

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PLoS One, 11*(6), e0157732. https://doi.org/10.1371/journal.pone.0157732.

Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. In *Paper presented at the proceedings of the 21st international conference on world wide web, Lyon, France*.

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

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (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.

March 25, 2007 Pontin, J. (2007). *Artificial intelligence, with help from the humans*. The New York Times http://www.nytimes.com/2007/2003/2025/business/yourmoney/2025Stream.html.

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

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior, 43*, 304–307. https://doi.org/10.1016/j.chb.2014.11.004.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740.

Siegel, J. T., Navarro, M. A., & Thomson, A. L. (2015). The impact of overtly listing eligibility requirements on MTurk: An investigation involving organ donation, recruitment scripts, and feelings of elevation. *Social Science & Medicine, 142*, 256–260. https://doi.org/10.1016/j.socscimed.2015.08.020.

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

Springer, V. A., Martini, P. J., Lindsey, S. C., & Vezich, I. S. (2016). Practice-based considerations for using multi-stage survey design to reach special populations on Amazon's Mechanical Turk. *Survey Practice, 9*.

Stanford Life-span Development Laboratory. (2016). *Future time perspective (FTP) scale*. Retrieved December, 2016, from https://lifespan.stanford.edu/projects/future-time-perspective-ftp-scale.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2017.06.007. online first.

Stewart, N., Ungemach, C., Harris, A. J. L. X., Bartels, D. M., Newell, B. R., Paolacci, G., et al. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment & Decision Making, 10*, 479–491.

Toma, C. L., Hancock, J. T., & Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin, 34*, 1023–1036. https://doi.org/10.1177/0146167208318067.

Wessling, K. S., Huber, J., & Netzer, O. (2017). MTurk character Misrepresentation: Assessment and solutions. *Journal of Consumer Research, 44*, 211–230. https://doi.org/10.1093/jcr/ucx053.

Wisconsin Department of Health Services. (2016). *The Myths about blindness and visual impairments*. Retrieved August 10, 2017, from https://www.dhs.wisconsin.gov/blind/adjustment/myths-blindvisual.htm.

Wolfe, C. R. (2017). SCiP: A discussion of surviving concepts and new methodologies. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0858-x. Epub online.