Journal of Behavioral Decision Making, *J. Behav. Dec. Making* (2012)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.1753

Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples

JOSEPH K. GOODMAN[1], CYNTHIA E. CRYDER[1]* and AMAR CHEEMA[2]
[1]*Olin Business School, Washington University in St. Louis, Saint Louis, MO, USA*
[2]*McIntire School of Commerce, University of Virginia, Charlottesville, VA, USA*

ABSTRACT

Mechanical Turk (MTurk), an online labor system run by Amazon.com, provides quick, easy, and inexpensive access to online research participants. As use of MTurk has grown, so have questions from behavioral researchers about its participants, reliability, and low compensation. In this article, we review recent research about MTurk and compare MTurk participants with community and student samples on a set of personality dimensions and classic decision-making biases. Across two studies, we find many similarities between MTurk participants and traditional samples, but we also find important differences. For instance, MTurk participants are less likely to pay attention to experimental materials, reducing statistical power. They are more likely to use the Internet to find answers, even with no incentive for correct responses. MTurk participants have attitudes about money that are different from a community sample's attitudes but similar to students' attitudes. Finally, MTurk participants are less extraverted and have lower self-esteem than other participants, presenting challenges for some research domains. Despite these differences, MTurk participants produce reliable results consistent with standard decision-making biases: they are present biased, risk-averse for gains, risk-seeking for losses, show delay/expedite asymmetries, and show the certainty effect—with almost no significant differences in effect sizes from other samples. We conclude that MTurk offers a highly valuable opportunity for data collection and recommend that researchers using MTurk (1) include screening questions that gauge attention and language comprehension; (2) avoid questions with factual answers; and (3) consider how individual differences in financial and social domains may influence results. Copyright © 2012 John Wiley & Sons, Ltd.

KEY WORDS research methods; surveys; sampling; online research; external validity

INTRODUCTION

> It is now possible for more people than ever to collaborate and compete in real time with more other people on more different kinds of work from more different corners of the planet and on a more equal footing than at any previous time in the history of the world (Friedman, 2005, p. 8).

The world's increased connectivity and access to information, called "flatness" by recent best-selling author Thomas Friedman, has changed not only business practices and social communication but also how academic researchers function. Connectivity allows new research methods and findings to be rapidly disseminated and adopted by researchers worldwide. Connectivity also provides unique advantages in human behavior research by providing more convenient access to broader groups of human participants than has been possible before, most notably through a recently developed online labor market called "Mechanical Turk" or "MTurk" (www.mturk.com).

MTurk, run by Amazon.com and named after an 18th century automated chess machine, provides an online workforce that allows people to complete work, or "Human Intelligence Tasks" (HITs), in exchange for money. Psychologists, behavioral economists, theoretical biologists, and consumer behavior researchers have recently started to recruit online study participants on MTurk. MTurk participants complete HITs, including academic studies, around the clock, allowing rapid data collection for as little as 10 cents per participant. MTurk is quickly being adopted by the research community; our own informal investigations of its use through MTurk postings and colleague surveys found that researchers from at least 16 of the top 30 US universities collect behavioral data via MTurk. However, at this time, MTurk is not yet widely accepted as a participant source.

A main concern about using MTurk for research is the belief that participants who are willing to participate in a study for only 10 cents must be unusual. Most importantly, they might be unusual in ways that challenge the validity of research investigations. Therefore, researchers in various fields have begun to look into data from MTurk participants to examine these concerns. These researchers have verified that MTurk demographic responses are accurate (Rand, 2011), validated the psychometric properties of MTurk responses (Buhrmester, Kwang, & Gosling, 2011), and replicated some of the classic findings in behavioral economics (Horton, Rand, & Zeckhauser, 2011; Suri & Watts, 2011) and decision-making research (Paolacci, Chandler, & Ipeirotis, 2010). Except for one study showing that MTurk workers were more risk-averse than non-MTurk participants (Paolacci et al., 2010) in the Asian Disease problem (Tversky & Kahneman, 1981), research has not identified significant differences between MTurk participants and traditional samples.

Although the body of evidence validating MTurk for use in behavioral research is growing, significant concerns remain. In particular, researchers and reviewers worry that MTurk workers do not pay sufficient attention to study materials. Additionally, researchers are concerned about the growing number of international participants on MTurk who may provide different responses because of language

*Correspondence to: Cynthia E. Cryder, Olin Business School, Washington University in St. Louis, One Brookings Dr., CB 1133, Saint Louis, MO 63130, USA. E-mail: cryder@wustl.edu

or cultural differences. Furthermore, MTurk workers, who participate in studies for extremely low amounts of money, may be especially peculiar in their attitudes about money and time, which are variables of particular interest to decision-making researchers.

In this paper, we directly compare MTurk samples with traditional student and community samples on several dimensions, including attention tests, a set of personality measures that include financial attitudes, and decision-making biases. Consistent with recent MTurk findings, we show that MTurk generally provides an excellent opportunity for inexpensive and efficient behavioral data collection with reliable results. However, we also find notable differences for MTurk participants that researchers should consider before using MTurk for their own research. We initially give a brief background about MTurk and the current state of behavioral research being conducted on MTurk.

MECHANICAL TURK

"MTurk" uses Internet crowd sourcing to connect potential workers with jobs or tasks, called "HITs." Both the tasks and the compensation are provided by the "requester" (an organization or individual), which pays MTurk "workers" based on the quality of their submitted work. At any moment, there are more than 100 000 HITs available to MTurk workers ranging in payment from $.01 to more than $10. The tasks vary widely in content and include, for example, conducting Web searches, editing audio transcripts, and completing questionnaires (see Mason & Suri (2012) for an excellent comprehensive guide about using MTurk). Workers participate in studies for multiple reasons, including supplemental income and enjoyment (Buhrmester et al., 2011; Paolacci et al., 2010). Paolacci et al. (2010) report that 61% of MTurk workers state that earning additional money motivates their work (though less than 14% state that it is a primary source of income) and 41% report completing HITs for entertainment.

MTurk workers are a diverse group of individuals in terms of location, with fewer than half (47%) living in the United States and about one-third living in India (Paolacci et al., 2010). Surveying MTurk workers ($n = 103$), we found that they learn about MTurk primarily from news articles and blogs (42%), friends (27%), and Internet searches (25%; only 3% learn about it from Amazon itself). Nevertheless, MTurk workers are not significant outliers in terms of their general demographics. Compared with the general US population, MTurk workers have a similar income distribution (with a slightly lower mean), are only slightly younger on average, have fewer children on average (though an average number for their age), and they tend to spend a day or less per week completing HITs (Ipeirotis, 2010).

Recently, behavioral researchers have dramatically increased their use of MTurk for data collection because of several advantages that MTurk offers. First, MTurk samples are inexpensive; a researcher can pay as little as 10–50 cents per participant for a short study. Second, MTurk allows for rapid data collection; often, an entire study can be completed within hours. Third, MTurk answers a call from some behavioral researchers for more representative samples compared with traditional student samples (e.g., Lynch, 1982; Parker & Fischoff, 2005; Peterson, 2001; Winer, 1999). MTurk reaches a more diverse population than even some community samples, allowing researchers to gain generalizability to broader populations (Buhrmester et al., 2011) and to test research questions across cultures (Eriksson & Simpson, 2010).

Because of the high interest in MTurk, recent research has attempted to systematically investigate MTurk's usefulness and reliability as a participant source. Rand (2011) found that demographic responses from MTurk were truthful and consistent; for example, more than 95% of MTurk participants reported their country location correctly as verified by IP address matching. Buhrmester et al. (2011) found that common fluctuations in compensation do not affect MTurk data quality. Buhrmester et al. (2011) also found that responses from MTurk participants were at least as reliable as those obtained from other non-MTurk samples and were also more representative of the general population compared with student participants and other online participants. Finally, MTurk participants exhibit similar judgment and decision biases (i.e., framing effects, the conjunction fallacy, and outcome bias) compared with students and online discussion board participants (Paolacci et al., 2010). Overall, these investigations conclude that MTurk is a high-quality source of behavioral participants.

RESEARCH QUESTIONS

Although research to date has widely endorsed MTurk, some pressing questions about using MTurk remain in the minds of both researchers and reviewers. Initially, do MTurk participants pay sufficient attention to study materials? A big concern is that MTurk participants are especially prone to skimming through study materials because they are unsupervised and poorly compensated. One study found high levels of attentiveness among MTurk participants (as measured by response to the question, "While watching the television, have you ever had a fatal heart attack?"; Paolacci et al., 2010). However, the answer to this screening question was easily recognizable, and almost all participants (95%), regardless of source, answered the question correctly. Because research has shown that unsupervised participants are less likely to pay attention to instructions (Oppenheimer, Meyvis, & Davidenko, 2009), we wished to further explore this issue by using more demanding attention checks. In both of our studies, we included a modified Instructional Manipulation Check (IMC; Oppenheimer et al., 2009) that was designed to assess whether or not participants were reading instructions carefully. We also examined whether English as a Second Language (ESL) speakers, a growing portion of MTurk participants, struggled with following these experiment instructions; demographic comparisons find that as many as one-third of MTurk participants are located in India alone, a proportion that is increasing with time (Paolacci et al., 2010). Language barriers

DOI: 10.1002/bdm

may compromise ESL participants' ability to comprehend and follow detailed directions.

A related question is whether MTurk participants have different cognitive capabilities compared with non-MTurk participants. This is an area of concern because cognitive predispositions have been found to be important moderators of decision patterns (Chatterjee, Heath, Milberg, & France, 2000; LeBoeuf & Shafir, 2003). Many MTurk participants (41%) cite enjoyment as a major reason for participating in MTurk studies (Paolacci et al., 2010), and therefore, it is possible that MTurk participants intrinsically enjoy cognitive activities more than community and student samples. Alternatively, college students, who by definition have chosen to pursue a higher education (arguably a cognitively rigorous endeavor), may be more likely to engage in effortful cognitive processing. To investigate these potential differences, we tested for differences in the Cognitive Reflection Test (CRT; Frederick, 2005) between MTurk participants and both student and community samples.

Although some research has examined the psychometric standards of MTurk participants' responses (Buhrmester et al., 2011), no work to date has examined whether personality features of MTurk participants differ from traditional student and community samples. To gain understanding of these dimensions, we administered standard personality inventories. Specifically, we administered the Ten-Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003), which captures the Big Five dimensions of personality (John & Srivistava, 1999). The Big Five personality dimensions comprise the most widely used and measured model of individual differences in psychology (Gosling et al., 2003). We also included a measure of global self-esteem (the single item self-esteem scale; Robins, Hendin, & Trzesniewski, 2001). Global self-esteem is also one of the most commonly studied constructs in psychology, predicting outcomes such as academic achievement, occupational success, and relationship health (Trzesniewski, Donnellan, & Robins, 2003).

Of special interest to decision-making and consumer behavior researchers are differences in the valuation and spending of money and time. Because MTurk participants are willing to complete Web tasks in exchange for little money, they are likely to differ on how they value money, material goods, and time. In both studies, we administered the time versus money scale (Cryder & Loewenstein, 2010) to measure how participants value time versus money, the Material Values Scale (MVS; Richins, 2004) to determine how they value material possessions, and the Tightwad-Spendthrift (TW-ST) scale (Rick, Cryder, & Loewenstein, 2008) to measure how painful it is for them to spend money.

Given the low compensation of MTurk participants, they might also respond unusually to decision tasks involving money and risk. Although previous research (Paolacci et al., 2010) has shown that MTurk participants exhibit some classic Judgment and Decision Making effects, specifically, framing effects (in the Asian Disease problem), the conjunction fallacy (Linda problem), and the outcome bias (physician problem), we do not yet know whether MTurk participants differ in terms of biases that deal with money and payoffs. In Study 2, we explored this proposition by testing for present bias and discounting asymmetries (Loewenstein, 1988; Malkoc & Zauberman, 2006; Thaler, 1981), risk aversion for gains, risk-seeking for losses, and the certainty effect (Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004; Kahneman & Tversky, 1979; Tversky & Kahneman, 1992).

Finally, MTurk also allows unique opportunities for its participants, who are completely unsupervised, to use the Internet as a resource. Although in some studies, this ability for participants to use the Internet may actually serve as a benefit, in others, it may serve as a hindrance, such as when participants can seek answers to factual study questions. We examine this issue using a classic judgment task that is designed to illustrate anchoring and adjustment and which relies on factual answers.

STUDY 1: COMPARING MTURK PARTICIPANTS WITH COMMUNITY PARTICIPANTS

In Study 1, we compared MTurk participants with a community sample from a large city in the northeastern United States. The study compared the two samples on several dimensions: individual differences, cognitive activity, native language, and an anchoring-and-adjustment judgment heuristic.

Participants and procedure

We sampled 107 MTurk participants, restricting participation to participants with an approval rate of at least 95% (i.e., 95% or more of that participant's previous submissions were approved by requesters; a 95% approval rate is the default MTurk cutoff). Participants received $.10 and were told that the study would take about 10 minutes. We posted a link for a Web-based study administered by Qualtrics, an online questionnaire software company, into an MTurk HIT. At the end of the questionnaire, participants received instructions to enter a unique code in the MTurk HIT to verify that they completed the study to receive payment. Although MTurk allows only one worker ID per person and the same ID may only do a HIT once, Qualtrics restrictions were also set to allow one response per IP address to provide additional protection against participants completing the study multiple times.

The community sample consisted of 60 participants in a commercial area of a middle class urban neighborhood. Participants were approached on the street by a research assistant and asked to participate in a short academic research study. They were paid $5 for a group of studies that included a paper-and-pencil version of this study. Although the two samples were paid significantly different amounts for the same study, the payments were intended to approximate the market rate for participation payment for each sample.

Participants in both samples completed identical survey measures in an identical order. To provide a comparison of basic individual differences between samples, participants initially answered questions from the TIPI (Gosling et al., 2003) that

DOI: 10.1002/bdm

measured the Big Five dimensions of personality (extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience; two traits per dimension, averaged, 1 = Disagree Strongly to 7 = Agree Strongly) and responded to a single-item self-esteem measure ("I have high self-esteem" 1 = Not very true of me to 5 = Very true of me; Robins et al., 2001). Participants then responded to three questions from the CRT to measure the extent to which the participant engaged in deliberate System 2 processing (Frederick, 2005). Next, participants answered several questionnaires related to money. Eight questions measured the extent to which participants value money versus time (Cryder & Loewenstein, 2010), four questions were from the TW-ST scale and measured the pain of paying (Rick et al., 2008), and nine questions were from the short version of the MVS (Richins, 2004).

Next, participants completed a classic anchoring and adjustment task that included a factual question (Oppenheimer, LeBoeuf, & Brewer, 2008; Tversky & Kahneman, 1974). Participants entered the last two digits of their phone number, indicated whether the number of countries on the continent of Africa is higher or lower than that number, and then estimated the number of countries in Africa.

Finally, participants answered demographic questions (age, gender, education, and native language) and finished with a modified IMC (Oppenheimer et al., 2009; see Appendix). The IMC gauged whether participants were paying careful attention to the instructions. For the IMC, participants read a paragraph about decision-making research, followed by a question that asked "what was this study about?" Participants were provided several multiple choice answers, including the obvious answer "your opinions and behaviors." However, the paragraph also included special instructions. These special instructions told participants to ignore the provided multiple choice answers, ". . .select the box marked 'other' and type 'decision making' on the line below." We note that in this study, the IMC included some answer choices that were obviously incorrect (such as "lions" and "tigers"). Such unusual options may have allowed some participants who were skimming questions and answers to notice that something was out of the ordinary, to pay more attention, and to successfully pass the IMC. This IMC design was intended to identify the group of participants paying the very least attention and to keep most participants in the sample. As will be seen in the results, this IMC still identified a significant number of participants who were below this (low) threshold of attention.

Results

Demographics

The MTurk sample did not significantly differ from the community sample in age ($M = 33$ for both groups), gender (female: MTurk = 58.8% versus Community = 51.7%) or education (modal and median education was 4-year bachelor's degree). More MTurk participants (27.5%) had ESL compared with the community sample (10.5%, $\chi^2(1,\ n = 159) = 5.79$, $p < .05$) presumably because of 25.5% of MTurk participants living outside the US (21% from India). As expected, ESL was correlated with residency ($r = .64$, $p < .001$): of non-US participants, 80.8% had ESL, whereas 9.8% of US participants had ESL. Sample characteristics are summarized in Table 1.

Attention and cognition

Our first research question asked whether MTurk participants paid as much attention to study materials as other participants. In this study, MTurk participants were just as likely to answer the IMC question correctly (81.1%) as the community sample (84.5%, $\chi^2(1,\ n = 167) < 1$), suggesting that the two samples were equally likely to read and follow instructions. Although non-US participants were less likely to correctly answer the IMC (70.4%) compared with US participants (88.5%, $\chi^2(1,\ n = 167) = 5.51$, $p < .05$), it seems that the difference is largely because of language comprehension: We found a significant difference between ESL and non-ESL participants in rates of answering the IMC correctly (non-ESL = 81% versus ESL = 61%, $\chi^2(1, n = 135) = 4.94$, $p < .05$). Our subsequent analyses include only participants who correctly answered the IMC; at the end of the results section, we examine whether our results change when including participants who failed the IMC and if the changes are similar when we exclude ESL or non-US participants.

A second question was whether MTurk participants, who are paid very little for their efforts, are as able to think through highly cognitive problems compared with other participants. Our results suggest that MTurk (M) CRT scores were not significantly different from the community (C) sample ($M_M = 1.17$ versus $M_C = .96$, $F(1, 133) = 1.10$, $p > .2$).

Money, time, and individual differences

Given that MTurk participants are willing to participate in studies for such small amounts of money, one might expect that they would value their time less and be tighter with their spending than would community participants. Consistent with this notion, MTurk participants valued their time slightly less and were more willing to participate in more tasks in exchange for money than were community participants. MTurk participants indicated they would participate in significantly more time-consuming tasks in exchange for money than would the community ($M_M = 5.13$ versus $M_C = 4.46$, $F(1, 132) = 5.22$, $p < .05$). MTurk participants also scored lower on the TW-ST scale, indicating that they are tighter with their spending than the community group ($M_M = 13.69$ versus $M_C = 15.86$, $F(1, 126) = 7.35$, $p < .01$). Finally, the MTurk group scored higher on the MVS ($M_M = 25.62$ versus $M_C = 23.27$, $F(1, 132) = 3.89$, $p = .05$), suggesting greater materialistic values than the community participants.

The results from the TIPI scale (Gosling et al., 2003) that measures the Big Five Personality factors showed that MTurk participants scored lower on extraversion ($M_M = 3.66$ versus $M_C = 4.56$, $F(1, 133) = 9.60$, $p < .01$) and emotional stability than community participants ($M_M = 4.23$ versus $M_C = 4.80$, $F(1, 133) = 4.88$, $p < .05$). Similarly, MTurk participants showed marginally lower self-esteem than

DOI: 10.1002/bdm

Table 1. Results—Study 1

	IMC	ESL**	Age	USA**	Female
MTurk	81.1%	27.5%	33.5	74.5%	58.8%
Community	84.5%	10.5%	32.9	100%	51.7%

	Cognitive reflection test		Time value of money**		Tightwad-spendthrift**		Material values*	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
MTurk	1.17	0.12	5.13	0.18	13.69	0.46	25.62	0.71
Community	0.96	0.16	4.46	0.23	15.86	0.66	23.27	0.95

	Extraversion**		Agreeableness		Conscientiousness		Emotional stability**		Openness to experience**		Self-esteem*	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE		
MTurk	3.66	0.17	4.81	0.13	5.33	0.14	4.23	0.16	4.98	0.12	3.3	0.12
Community	4.56	0.23	5.09	0.17	5.52	0.18	4.80	0.21	5.77	0.15	3.67	0.15

**$p < .05$; MTurk participants are statistically different from community participants.
*$p < .06$; MTurk participants are marginally different from community participants.
[i]$p < .05$; MTurk × IMC interaction.
IMC, Instructional Manipulation Check; ESL, English as a Second Language; SE, standard error.

did the community sample ($M_M = 3.30$ versus $M_C = 3.67$, $F(1, 133) = 3.92$, $p = .055$). Despite being active users of new technologies, MTurk participants also scored lower on openness to experience ($M_M = 4.98$ versus $M_C = 5.77$, $F(1, 133) = 16.98$, $p < .001$). There was no significant difference between samples on the dimensions of agreeableness and conscientiousness.

Anchoring and adjustment

To examine the anchoring and adjustment data, we regressed participants' country number estimates on their phone number anchors (mean-centered), sample group, and their interaction. We found that the larger the participants' phone number, the greater the participants' estimate of the number of countries in Africa (intercept = 41.37, $\beta = .19$, $t(129) = 3.20$, $p < .01$). However, we also found a significant interaction ($\beta = -.13$, $t(129) = 3.03$, $p < .01$) indicating that although the community showed a significant anchoring and adjustment effect ($\beta = .45$, $t(129) = 4.31$, $p < .01$), MTurk participants did not ($\beta = .06$, $t(129) < 1$).

Further analysis revealed that MTurk participants might be using their freedom from experimenter monitoring to check factual answers on the Internet. Ten percent of MTurk respondents "estimated" the correct number of countries in Africa as reported by a quick Google search (53 countries) or Wikipedia query (54 countries), whereas none of the community participants gave these precisely correct responses ($\chi^2(1, n = 156) = 5.98$, $p < .02$).

Because this tendency to fact-check is important for many types of investigations, we ran two short follow-up anchoring and adjustment studies to better understand when fact-checking is likely to be a problem. One potential reason for looking up answers is that participants feel pressure to give correct answers to receive their participation payments. To test this proposition, we ran the same anchoring and adjustment study on MTurk while telling half the participants that they would get paid regardless of their answer. This intervention did not significantly affect the rate of correct answers (14.3%) compared with a control condition (21%, $\chi^2(1, n = 94) < 1$).

To further examine the factors that influence cheating incidence, we ran a second study that manipulated (1) compensation and (2) whether or not we told participants the following: "Please do NOT use external sources like the Internet to search for the correct answer." In a 3 (compensation: none, \$.10, \$1) × 2 (instructions: control versus do not search) between subject factorial design experiment, we found main effects for compensation and instructions (with no significant interactions; see Table 2 for percentages). First, participants who were paid for correct answers were significantly more likely to look up a correct answer ($\chi^2(1, n = 270) = 17.21$, $p < .001$), and the amount of payment had only a marginal effect (\$.1 = 36.2% versus \$1 = 49.6%, $\chi^2(1, n = 270) = 3.03$, $p = .08$). Second, simply asking MTurk participants to not use the Internet to search for the correct answer was sufficient to significantly reduce cheating from 40.1% to 27.2% ($\chi^2(1, n = 270) = 5.76$, $p < .05$), although it did not eliminate cheating entirely.

Instructional Manipulation Check, English as a second language, and US participants

The final issue we addressed was how including participants who failed the IMC affected our results, as well as whether ESL and non-US participants showed different patterns. The goal of the IMC is to identify participants who are not following instructions, potentially increasing Type II error. Our results showed that few conclusions change when we include all participants, though not surprisingly, statistical

Table 2. Results—Study 1 follow-up: percentage of participants "estimating" the correct number of countries in Africa (i.e., percentage searching online)

	No payment	\$.10	\$1
Control	23%	38%	60%
Additional "do not check" instruction	8%	35%	39%

DOI: 10.1002/bdm

significance suffers. We found only one significant MTurk by IMC interaction (on emotional stability, $F(1, 159) = 4.24$, $p < .05$). As an even more conservative test, we also examined whether any of our significant results became nonsignificant, or reversed, when those who failed the IMC were included in the analysis. Of our significant findings, two became nonsignificant when we included those who failed the IMC: material values ($F(1, 157) = 2.32$, $p = .13$) and emotional stability ($F(1, 164) = 2.09$, $p = .15$). None of our effects showed a significant opposite result. In sum, the results show that including participants who incorrectly answered the IMC adds statistical noise and potentially increases the incidence of Type II error.

Given the growing number of international MTurk participants, we also examined whether participants with ESL and non-US participants differed significantly from the rest of our sample. As previously noted, ESL and non-US participants were more likely to fail the IMC, but the correlations were modest ($r_{\text{ESL, IMCfail}} = .18$, $p < .05$; $r_{\text{non-US, IMCfail}} = .16$, $p = .05$). Examining only participants who passed the IMC, we did not find any significant MTurk by ESL interactions ($ps > .5$). Examining all participants, we found only one significant MTurk by ESL interaction, on emotional stability, in which non-ESL participants showed a larger difference than the ESL participants ($F(1, 154) = 4.23$, $p < .05$). Limiting our analysis to only non-ESL participants, we find that one of our significant results became nonsignificant (material values, $F(1, 121) = 1.43$, $p > .2$). We find similar results when we limit our analysis to only US participants: two of our significant results became nonsignificant (material values, $F(1, 129) < 1$, and emotional stability, $F(1, 130) = 2.62$, $p > .1$). In sum, ESL and non-US participants are more likely to fail the IMC, but more significant effects emerge when we filter by IMC instead of by ESL or by location.

Discussion

In Study 1, we found several similarities between MTurk and community participants. Initially, the two groups were similar in gender and age and similar in cognitive effort and ability as measured by the CRT, and according to performance on the IMC, they were equally likely to follow instructions. Recall that we used a 95% worker approval rating on MTurk as a prerequisite for a worker to complete our survey. This is the default cutoff level used on MTurk. We speculate that using a lower cutoff (or no cutoff) would have led to a larger proportion of MTurk participants failing the IMC, but we do not have empirical evidence to test this hypothesis.

There were important differences between groups as well. MTurk participants differed in terms of their valuation of money, time, and material goods. MTurk participants valued money more than time compared with the community sample, and they were more likely to be self-reported tightwads (versus spendthrifts). They also scored higher on the MVS, which measures materialism and the meaning of material goods in participants' lives.

Interestingly, a coherent pattern emerged on the basic individual difference measures. MTurk participants were less extraverted and more socially unstable than the community participants. MTurk participants also showed slightly lower self-esteem, which is related to extraversion (Robins et al., 2001) and relationship health (Trzesniewski et al., 2003). Although not predicted, these results could possibly be explained by (a) the demographic diversity of the sample; (b) something unique about MTurk participants, such as a general predilection toward introversion and other related dispositional constructs traits; and/or (c) something unique about the community sample, such as the potential extraverted nature a group of individuals who complied with a request from a stranger on the street to participate in an experiment. In Study 2, we further investigate these basic personality characteristics of the MTurk sample to determine whether the observed differences are more likely because of the unique nature MTurk participants or community participants.

Finally, we also found a difference in terms of heuristic use: MTurk participants did not seem to rely upon a prior anchor when estimating the number of countries in Africa. Instead, it appears that at least some MTurk respondents were looking up the correct answer online. Follow-up studies indicated that even small incentives amplify the tendency to cheat and that requests to not look up answers significantly decrease, but do not entirely eliminate, cheating.

STUDY 2: COMPARING MTURK PARTICIPANTS WITH STUDENT PARTICIPANTS

Our main goal in Study 2 was to compare MTurk participants to a traditional undergraduate student research pool, and while doing so, gain a greater understanding of the unique features of MTurk respondents. In this study, we also wanted to control for survey format (paper and pencil versus computer based). It is possible that differences between the MTurk participants and the community participants in the previous study were influenced by differences in questionnaire format—the community sample used paper and pencil instruments, whereas the MTurk sample used a computer.

Participants and procedure

We sampled 207 participants from MTurk, again restricting the sample to participants with an approval rate of 95% or higher. Participants received $.20 for completing the study and were told that the HIT takes approximately 15 minutes to complete. We posted a link to the Qualtrics questionnaire in the HIT, and after completing the questionnaire, participants were instructed to return to the HIT and enter their unique code. Actual average time of completion (excluding five outliers who each took over 1 hour) was 16 minutes 35 seconds. We restricted the HIT so that it could only be completed once per MTurk ID, and we also restricted Qualtrics to accept only one response per IP address.

We sampled 131 students from a Midwestern university in the United States. Participants received credit in their introductory business courses for participation in a 30-minute session that included several studies. Half of the students were

DOI: 10.1002/bdm

randomly placed in a computerized condition and half in an equivalent paper-and-pencil condition. Average time of completion on the computer was 12 minutes 41 seconds, and with paper and pencil, 14 minutes 21 seconds.

To further examine differences in decision making related to money and payoffs, participants initially responded to four risky choice prospects, adapted from previous research, testing risk aversion for gains, risk-seeking for losses, and the certainty effect (Barron & Erev, 2003; Hertwig et al., 2004; Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). The four problems are included in Table 3. Participants then responded to a discounting task adapted from previous literature testing present bias (Thaler, 1981) and the delay/expedite asymmetry (Loewenstein, 1988; Malkoc & Zauberman, 2006).

Then, as in Study 1, participants completed the CRT (Frederick, 2005), the 10-item inventory for the Big Five personality dimensions (TIPI; Gosling et al., 2003), the single-item self-esteem scale (Robins et al., 2001), the time versus money scale (Cryder & Loewenstein, 2010), the TW-ST (Rick et al., 2008), and the MVS (Richins, 2004). New to this study, participants also completed the Maximizer/Satisficer Scale (Nenkov, Morrin, Ward, Schwartz, & Hulland, 2008) to examine cognitive predispositions and preference for optimal outcomes at the expense of extra effort. Finally, participants completed the demographic questions from Study 1 and a new, more conservative version of the IMC (Oppenheimer et al., 2009) to test whether participants were closely reading the instructions of the experiment. This version of the IMC required more careful reading to give a "correct" answer than the version used in Study 1 by including reasonable, but incorrect, answer choices that were less likely to trigger suspicion (see Appendix).

Results

We created two orthogonal contrast codes to capture the nested design of the study (Rosenthal, Rosnow, & Rubin, 2000). The first code compared the MTurk participants with the student group, and the second code compared the computerized and paper and pencil participants within the student group. The second contrast code (computer versus paper and pencil) was generally not significant; thus, we report the data collapsed across conditions noting any differences that approach significance (i.e., $p < .1$) in Table 4.

Demographics

Not surprisingly, MTurk (M) participants were significantly older than the student (S) participants ($M_M = 31.0$ versus $M_S = 19.4$, $F(1, 334) = 164$, $p < .001$), and the median level of education was higher (Mdn_M = "bachelor's degree") than the students (Mdn_S = "some college", Wilcoxon $Z = 10.52$, p .001). As in Study 1, more MTurk participants had ESL (57.0%) compared with the student sample (20.3%, $\chi^2(1, n = 335) = 57.01$, $p < .001$), and 62.8% of MTurk participants were living outside the US (52.2% from India). As in Study 1, ESL was correlated with residency ($r_{ESL,\ non\text{-}US} = .82$, $p < .001$): of non-US participants, 87.8% had ESL, whereas 9.4% of US participants had ESL. Sample characteristics are summarized in Table 4.

Cognition and attention

The IMC tests whether participants were paying attention and following instructions. In Study 2, we used a significantly more difficult IMC test than in Study 1 to see if increased difficulty mattered. We found that MTurk participants were significantly less likely than student participants to correctly answer the more difficult IMC question ($M_M =$ 66.2% versus $M_S = 88.5\%$, $\chi^2(1, n = 338) = 19.51$, $p < .001$). We also again found a significant difference in rates of answering the IMC correctly between non-ESL (71%) and ESL participants (29%, $\chi^2(1, n = 335) = 54.27$, $p < .001$). As in Study 1, our subsequent analyses only consider those participants who followed instructions on the IMC. At the end of this results section, we again discuss how our results change when we include those who failed the IMC and if the changes are similar when we exclude ESL or non-US participants.

Students scored significantly higher on the CRT than did the MTurk participants ($M_M = 1.23$ versus $M_S = 1.70$, $F(1, 250) = 10.49$, $p < .01$). Note that there was no difference between MTurk participants and the community sample in Study 1, suggesting that it may be student participants who are unique in terms of their reflective thought. Despite differences between students and MTurk participants in deliberative thought on the CRT, we did not find any differences in terms of preference for optimal outcomes as measured by the Maximizer/Satisficer Scale ($M_M = 37.36$ versus $M_S = 38.72$, $F(1, 250) = 2.25$, $p > .1$).

Table 3. Risk prospects—Study 2

	Alternative 1		Alternative 2		*P*(R)		
	$, Probability	E(X)	$, Probability	E(X)	MTurk	Student computer	Student paper
Prospect 1	$4, .8	$3.20	<u>$3, 1.0</u>	$3.00	15%	21%	15%
Prospect 2	$320, .8	$256.00	<u>$240, 1.0</u>	$240.00	14%*	21%	25%
Prospect 3	<u>$4, .2</u>	$0.80	$3, .25	$0.75	48%**	70%	81%
Prospect 4	<u>−$3, 1.0</u>	−$3.00	−$4, .8	−$3.20	76%	82%	76%

*$p < .06$; MTurk < students.
**$p < .001$; MTurk < students.
Underlined choice predicted using Prospect Theory. E(X), expected value of alternative; *P*(R), percentage selecting alternative 1.

DOI: 10.1002/bdm

Table 4. Results—Study 2

	IMC**	ESL**	USA**	Female**	Some college**	Current student**	Age**	Maximizer/ Satisficer	
							Mean	Mean	SE
MTurk	66.2%	56.3%	37.2%	42.5%	92.3%	23.7%	31.0	37.37	0.61
Student computer	86.4%	9.4%	100%	53.9%	100.0%	100%	19.5	38.79	0.95
Student paper	90.8%	20.0%	100%	59.4%	100.0%	100%	19.3	38.64	0.93

	Cognitive Reflection Test**†		Time versus money		Tightwad-spendthrift~		Material values	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
MTurk	1.23	0.1	5.49	0.12	14.56	0.34	27.57	0.54
Student computer	1.68	0.15	5.26	0.19	14.14	0.52	27.84	0.83
Student paper	1.71	0.15	5.27	0.19	13.96	0.52	27.52	0.83

	Extraversion**		Agreeableness[b]		Conscientiousness**		Emotional Stability**		Openness to Experience		Self-Esteem**	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
MTurk	4.03	0.12	5.06	0.11	5.18	0.10	4.47	0.12	5.08	0.09	3.46	0.08
Student computer	4.97	0.19	4.66	0.16	5.43	0.16	4.98	0.18	5.20	0.14	3.77	0.13
Student paper	4.69	0.19	5.09	0.16	5.58	0.16	4.66	0.18	5.32	0.14	3.72	0.13

**$p < .05$; MTurk participants are significantly different from student participants.
[b]$p < .06$; student computer participants are marginally different from student paper participants.
[i]$p < .05$; MTurk versus students × IMC interaction.
~$p < .1$; MTurk versus students × IMC interaction.
†$p < .05$; MTurk versus students × ESL interaction.
IMC, Instructional Manipulation Check; ESL, English as a Second Language; SE, standard error.

Money, time, and individual differences

In terms of basic individual differences, we see results similar to Study 1. Most notably and consistent with Study 1, MTurk participants were less extraverted ($M_M = 4.03$ versus $M_S = 4.83$, $F(1, 250) = 19.75$, $p < .001$), less emotionally stable ($M_M = 4.47$ versus $M_S = 4.82$, $F(1, 250) = 4.17$, $p < .05$), and had less self-esteem ($M_M = 3.46$ versus $M_S = 3.75$, $F(1, 250) = 5.18$, $p < .05$) than student participants. Differing from Study 1, MTurk participants were also less conscientious than the student participants ($M_M = 5.18$ versus $M_S = 5.50$, $F(1, 250) = 4.43$, $p < .05$) and did not show a significant difference in their openness to new experiences ($M_M = 5.08$ versus $M_S = 5.26$, $F(1, 250) = 1.81$, $p > .1$).

Although Study 1 found that MTurk participants valued money (versus time) more than the community, Study 2 found that MTurk participants did not value their money versus time differently than student participants ($M_M = 5.49$ versus $M_S = 5.28$, $F(1, 250) = 1.52$, $p > .2$). Furthermore, the MTurk group was just as tightwad as students ($M_M = 14.56$, $M_S = 14.05$, $F(1, 248) = 1.02$, $p > .2$) and showed no differences from students in material values ($M_M = 27.57$, $M_S = 27.69$, $F(1, 249) < 1$). These similarities between students and MTurk participants on spending constructs could be because both students and MTurk participants live with constrained budgets.

Given the low compensation of MTurk participants, a goal of Study 2 was to examine whether MTurk participants differed from undergraduates on decision tasks related to money and risk, specifically, risk aversion for gains, risk-seeking for losses, the certainty effect (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), present bias (Thaler, 1981), and delay/expedite asymmetries (Loewenstein, 1988). Both student and MTurk participants were risk-averse for gains for both small ($P(R)_1$: MTurk = .15, Student = .18) and large gambles ($P(R)_2$: MTurk = .14, Student = .28), and they were risk-seeking for losses ($P(R)_4$: MTurk = .76, Student = .79). Both groups exhibited the classic certainty effect ($P(R)_1$ MTurk = .15 versus $P(R)_3$ MTurk = .48, $F(1, 136) = 50.75$, $p < .001$; $P(R)_1$ Student = .18 versus $P(R)_3$ Student = .76, $F(1, 115) = 116.79$, $p < .001$). Our results suggest that Prospect Theory patterns of risk taking do apply to both MTurk and student participants, although we did find some minor differences. MTurk participants were marginally more risk-averse for one of the four prospects (Prospect 2) than the student participants ($\chi^2(1, n = 253) = 3.58$, $p = .058$), and the student sample showed a significantly stronger certainty effect ($F(1, 250) = 130.27$, $p < .001$).

Time preference results show that although both groups showed statistically significant present bias in a delay frame ($M_{3months} = 4.50$ versus $M_{12months} = 2.06$, $F(1, 233) = 217.86$, $p < .001$) and expedite frame ($M_{3months} = .38$ versus $M_{12months} = .20$, $F(1, 230) = 28.73$, $p < .001$), MTurk participants exhibited significantly stronger tendencies toward present bias in both frames (Delay: $F(1, 233) = 4.24$, $p < .05$; Expedite: $F(1, 230) = 5.86$, $p < .05$).[1] The two groups exhibited delay/expedite asymmetries at similar rates: present bias was greater in the delay frame compared with the expedite frame

[1]We used continuously compounded discount rates as the dependent variable (Malkoc & Zauberman, 2006; Thaler, 1981). We removed from the analysis 16 participants who indicated a negative discount rate and did not understand the instructions (they were willing to pay more than $15 to expedite a $15 payment).

DOI: 10.1002/bdm

($F(1, 229) = 166.67$, $p < .001$), and discount rates were significantly greater when delaying 3 months versus expediting 3 months ($F(1, 232) = 198.53$, $p < .001$) and when delaying 12 months versus expediting 12 months ($F(1, 229) = 145.20$, $p < .001$). These comparisons are listed in Table 5.

Instructional Manipulation Check, English as a second language, and US participants

We further investigated the IMC to examine how our results would change if we included those who failed the IMC. There was only one marginally significant MTurk by IMC interaction ($F(1, 330) = 3.44$, $p < .1$): the difference between MTurk participants and students on the TW-ST scale was greater for those who failed the IMC than those who passed. As in Study 1, our main conclusions do not change when we include participants who failed the IMC in the analyses, but statistical significance suffers. Of the significant findings that we found, differences in emotional stability become marginally significant ($F(1, 335) = 3.29$, $p = .07$) when we include those who failed the IMC.

Examining ESL and non-US participants, we again found that IMC failures are positively correlated with ESL and non-US location ($r_{ESL,\ IMCfail} = .40$, $p < .001$; $r_{non\text{-}US,\ IMCFail} = .38$, $p < .001$). We found one significant MTurk by ESL interaction (CRT, $F(1, 329) = 3.90$, $p < .05$): the difference between MTurk and student participants on the CRT was smaller for ESL participants compared with non-ESL. More importantly, we found that statistical significance suffered when we filtered by ESL and only examined non-ESL participants; two of our significant findings became marginally significant or nonsignificant: emotional stability ($F(1, 197) < 1$) and conscientiousness ($F(1, 197) = 3.72$, $p = .06$). We found similar results when we limited our analysis to non-US participants; emotional stability ($F(1, 205) = 1.88$, $p > .15$) and conscientiousness ($F(1, 205) = 1.61$, $p > .2$) became nonsignificant. In sum, ESL and non-US participants were more likely to fail the IMC, but the results suggest that it is more efficient to filter by IMC because it filters fewer participants and allows us to detect more significant differences.

Table 5. Difference in discount rates and delay/expedite asymmetries—Study 2

	Delay 3 versus delay 12 months	Expedite 3 versus expedite 12 months	Delay 3 versus expedite 3 months	Delay 12 versus expedite 12 months
MTurk	2.8**	0.25**	4.45	1.99
Student computer	2.51[b]	0.07	4.46	2.03
Student paper	1.7	0.09	3.14	1.52
Overall	2.57	0.19	4.23	1.92

All differences are significantly different from zero, $p < .05$.
**$p < .05$; different from the student sample.
[b]$p < .1$; student computer participants are marginally different from student paper participants.

Discussion

In many ways, MTurk participants were similar to the student sample; they had similar attitudes about money and exhibited the same classic decision-making biases. However, MTurk participants also showed important differences compared with the student participants: MTurk participants were less likely to correctly answer the IMC, they scored lower on the CRT, and they were less extraverted with lower emotional stability and self-esteem. Although MTurk participants were slightly more risk-averse and showed a stronger certainty effect than the student population (they were more likely than student participants to prefer payoffs with certainty compared with gambles with higher expected values), in general, MTurk participants exhibited the same classic decision-making biases. MTurk participants were risk-averse for gains, risk-seeking for losses, susceptible to the certainty effect, present biased, and exhibited delay/expedite asymmetries.

GENERAL DISCUSSION

In two studies, we compared participants from a new and increasingly popular online labor market, MTurk, with participants from traditional community and student samples. Our results not only showed many similarities between MTurk and these traditional samples, but we also identified important dimensions on which MTurk participants differed.

First, in our IMC attention test (Oppenheimer et al., 2009), which required careful reading of study materials and proficient English comprehension, our results showed that MTurk participants performed more poorly compared with student participants. It is important to note, however, that participants in all groups failed at a sizeable rate. More importantly, we found that simply administering the attention test and filtering participants by whether they correctly answered the IMC or not reduced statistical noise; including participants who failed the IMC reduced the likelihood of finding statistically significant differences between groups. Although we found that IMC failure was correlated with a participant being from outside the US and ESL, the results suggest the IMC was the most efficient filter: the IMC excluded fewer people and uncovered more statistically significant differences.

Second, in an anchoring and adjustment task in Study 1, we found a significant number of MTurk participants who "estimated" the precisely correct number of countries in Africa. No community participants estimated these precisely correct answers, suggesting that MTurk participants are using their freedom from supervision to check answers online. The good news for researchers is that there are easy ways to reduce this "cheating." A follow-up study revealed that although small incentives significantly increased this tendency to cheat, simply asking participants to not look up answers significantly reduced correct estimates. These findings fit with the cheating behavior literature showing that subtle situational variations significantly influence cheating rates (e.g., Mazar, Amir, & Ariely, 2008; Shu, Gino, & Bazerman, 2011).

DOI: 10.1002/bdm

Third, we found that MTurk participants were significantly and consistently different on basic personality variables, including some Big Five personality dimensions (Gosling et al., 2003; John & Srivistava, 1999), and they also showed some differences on dimensions related to money and payoffs. Compared with non-MTurk participants, MTurk participants were less extraverted, less emotionally stable, and had lower self-esteem. MTurk participants also exhibited attitudes about money and time that were more similar to student participants than to community participants. MTurk participants valued money more than time compared with the community, were more likely to be "tightwads," and valued material possessions more highly than the community. Compared with students, MTurk participants did not differ on these dimensions. It seems that MTurk participants may be similar to students in terms of their financial outlook.

It is important to note that we found many commonalities between MTurk participants and our traditional samples, contributing to the growing literature showing that MTurk participants give responses similar to other traditionally used samples (Buhrmester et al., 2011; Paolacci et al., 2010; Rand, 2011). MTurk participants were present biased, showed delay/expedite asymmetries, were risk-averse for gains, risk-seeking for losses, and showed the certainty effect—all with almost no significant differences in effect sizes from other samples.

Recommendations

On the basis of the findings suggesting similarities between MTurk and traditional samples and the previous research showing the benefits and reliability of MTurk (Buhrmester et al., 2011; Mason & Suri, 2012; Paolacci et al., 2010; Rand, 2011), we highly recommend MTurk to behavioral decision-making researchers because of its reliability, low cost, speed of data collection, and heterogeneity of participants. The benefits notwithstanding, our results also suggest that important unique features of MTurk participants should be considered before selecting MTurk as a participant source. We recommend that researchers using MTurk take note of several factors.

Initially, our findings suggest that an attention check is likely to help improve statistical power in all types of studies and reduce Type II error, but it may be especially useful with unsupervised samples such as MTurk. We caution researchers when using MTurk for studies that require participants to pay careful attention to study materials and instructions. For example, MTurk may not be appropriate for long or complicated studies in which participants may be more likely to lose attention and not follow instructions. Previous research (Paolacci et al., 2010) found that MTurk participants were equally attentive as other participants when the study was short (~5 minutes) and the attention check question had an easily identifiable answer. In our research, in which the studies were longer (~16 minutes) and the attention check question required careful reading, MTurk participants performed significantly worse. Following prior research, we placed the attention check at the end of the survey and this may have increased the likelihood that participants failed the check because of fatigue. We speculate that placing the check earlier in the survey would decrease failure rates for both MTurk and traditional participants, but it is possible that MTurk participants were more susceptible to fatigue compared with the students.

Our results from the CRT also suggest that MTurk participants may not be as motivated as student participants to engage in deliberate System 2 cognitive processing. Future research should continue to examine the boundaries of MTurk participants' attention and concentration. One way to mitigate the lower attention associated with MTurk samples could be to emphasize the scientific importance of the study to participants and encourage them to be attentive. Although laboratory studies do not always employ these measures, the presence of experimenters in a university setting and/or more complicated procedures may automatically have that effect in the lab, which is difficult to replicate online.

We also recommend caution when researchers ask MTurk participants questions with factual answers, unless one wants to test their Internet searching abilities. Although most behavioral study questions are not factual in nature, there are many examples that are, such as the questions commonly used in the anchoring and adjustment task (Tversky & Kahneman, 1974) in which we originally uncovered the fact-checking problem, as well as those that measure competencies, for example, in the CRT (Frederick, 2005). When using such factual knowledge questions is unavoidable, researchers should consider using simple interventions that ask participants not to look up answers, as our results showed that they are effective in reducing the rate at which MTurk participants searched for correct answers.

Finally, we suggest that researchers carefully consider how individual differences may influence their investigations. Specifically, researchers should take note when examining and interpreting results from MTurk participants that are related to extraversion and self-esteem—for which MTurk workers scored consistently lower—and money and spending—for which MTurk workers' attitudes were similar to students but different from a community sample.

Conclusions

Recent research about the use of MTurk for behavioral research has concluded that MTurk has many benefits, making it suitable for a wide range of behavioral research. We agree—we found that MTurk participants produced reliable results that are consistent with previous decision-making research. However, we also found important differences between MTurk participants and community and student participants. To mitigate concerns that may arise from these differences, we recommend that researchers use screening procedures to measure participants' attention levels and take into account that MTurk participants may vary from non-MTurk participants on social and financial traits.

DOI: 10.1002/bdm

APPENDIX. WORDING OF THE INSTRUCTIONAL MANIPULATION CHECK QUESTION:

Study 1

Research in decision making shows that people, when making decisions and answering questions, prefer not to pay attention and minimize their effort as much as possible. Some studies show that over 50% of people don't carefully read questions. If you are reading this question and have read all the other questions, please select the box marked 'other' and type 'Decision Making' in the box below. Do not select "predictions of your own behavior." Thank you for participating and taking the time to read through the questions carefully!

What was this study about?

A Predictions of your own behavior
B Lions
C Tigers
D Other _____

Study 2

Research in decision making shows that people, when making decisions and answering questions, prefer not to pay attention and minimize their effort as much as possible. Some studies show that over 50% of people don't carefully read questions. If you are reading this question and have read all the other questions, please select the box marked 'other' and type 'Decision Making' in the box below. Do not select "predictions of your own behavior." Thank you for participating and taking the time to read through the questions carefully!

What was this study about?

A Predictions of your own behavior
B Predictions of your friends' behavior
C Political preferences
D Other _____

ACKNOWLEDGEMENTS

The first two authors contributed equally to this manuscript. We thank Selin Malkoc, Reade Alexander, Josh Morris, Matt Williams, Derek Wilton, Lab Manager Rachel London, and the entire CB Research Lab for research assistance, and the Center for Behavioral Decision Research at Carnegie Mellon University for assistance in data collection.

REFERENCES

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*, 215–233.

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

Chatterjee, S., Heath, T. B., Milberg, S. J., & France, K. R. (2000). The differential processing of price in gains and losses: The effects of frame and need for cognition. *Journal of Behavioral Decision Making*, *13*, 61–75.

Cryder, C., & Loewenstein, G. (2010). The time versus money scale. Unpublished data, Olin Business School, Washington University in St. Louis.

Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, *5*, 159–163.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25–42.

Friedman, T. L. (2005). *The world is flat: A brief history of the twenty-first century*. New York, NY: Farrar, Straus, and Giroux.

Gosling, S. D., Rentfrow, P. J., & Swann Jr., W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, *37*, 504–528.

Hertwig, R., Barron G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *8*, 534–539.

Horton J. J., Rand D. G., & Zeckhauser R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*, 399–425.

Ipeirotis, P. (2010). Demographics of Mechanical Turk. (CeDER Working Paper-10-01). New York University. Retrieved from http://hdl.handle.net/2451/29585.

John, O. P., & Srivistava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York/London: Guilford Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292.

LeBoeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making*, *16*, 77–92.

Loewenstein, G. (1988). Frames of mind in intertemporal choice. *Management Science*, *34*, 200–214.

Lynch Jr., J. G. (1982). On the external validity of experiments in consumer research. *Journal of Consumer Research*, *9*, 225–239.

Malkoc, S. A., & Zauberman G. (2006). Deferring versus expediting consumption: The effect of outcome concreteness on sensitivity to time horizon. *Journal of Marketing Research*, *43*, 618–627.

Mason, W. & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavioral Research Methods*, *44*, 1–23.

Mazar, N., Amir, O., Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*, 633–644.

Nenkov, G., Morrin, M., Ward, A., Schwartz, B., & Hulland J. (2008). A short form of the maximization scale: Factor structure, reliability, and validity studies. *Judgment and Decision Making*, *3*, 371–388.

Oppenheimer, D., Meyvis T., & Davidenko N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872.

Oppenheimer, D. M., LeBoeuf, R. A., & Brewer, N. T. (2008). Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming. *Cognition*, *106*, 13–26.

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

Parker, A., & Fischoff, B. (2005). Decision-making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making*, *18*, 1–27.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 450–461.

Rand, D. G. (2011). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, DOI: 10.1016/j.jtbi.2011.03.004.

Richins, M. L. (2004). The material values scale: Measurement properties and development of a short form. *Journal of Consumer Research*, *31*, 209–219.

DOI: 10.1002/bdm

Rick, S. I., Cryder, C. E., & Loewenstein, G. (2008). Tightwads and spendthrifts. *Journal of Consumer Research*, *34*, 767–782.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*, 151–161.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge: Cambridge University Press.

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin*, *37*, 330–349.

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in Web-based, networked public goods experiments. *PLoS One*, *6*(3), 1–8.

Thaler, R. H. (1981). Some empirical evidence on dynamic inconsistency. *Economic Letters*, *8*, 201–207.

Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2003). Stability of self-esteem across the life-span. *Journal of Personality and Social Psychology*, *84*, 205–220.

Tversky, A., & Kahneman D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *211*, 453–458.

Tversky, A., & Kahneman D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of Uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Winer, R. S. (1999). Experimentation in the 21st century: The importance of external validity. *Journal of the Academy of Marketing Science*, *27*, 349–358.

Authors' biographies:

Joseph K. Goodman is an Assistant Professor of Marketing, Olin Business School, Washington University in St. Louis. He earned his PhD in marketing from The University of Texas at Austin. His research interests are in consumer variety seeking and choice overload, and how consumer experiences and superstitions influence happiness.

Cynthia Cryder is an Assistant Professor of Marketing, Olin Business School, Washington University in St. Louis. She earned her PhD in Behavioral Decision Research from Carnegie Mellon. Her research focuses on altruistic and financial decisions and has been published in journals including *Psychological Science* and the *Journal of Consumer Research*.

Amar Cheema is an Associate Professor of Marketing, McIntire School of Commerce, University of Virginia. He earned his PhD in marketing from the University of Colorado at Boulder. His research interests include behavioral decision theory, pricing and promotion effects, auctions and online purchase behavior, and word-of-mouth influences.

Authors' addresses:

Joseph K. Goodman and Cynthia E. Cryder, Olin Business School, Washington University in St. Louis, Saint Louis, MO, USA.

Amar Cheema, McIntire School of Commerce, University of Virginia, Charlottesville, VA, USA.

DOI: 10.1002/bdm