Computers in Human Behavior 43 (2015) 304–307

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Research Report

A reliability analysis of Mechanical Turk data

Steven V. Rouse

Social Sciences Division, Pepperdine University, Malibu, CA 90263, United States

ARTICLE INFO

Article history:

Keyword:
Psychometrics

ABSTRACT

Amazon's Mechanical Turk (MTurk) provides researchers with access to a diverse set of people who can serve as research participants, making the process of data collection a streamlined and cost-effective one. While a small number of studies are often cited to support the use of this methodology, there remains a need for additional analyses of the quality of the research data. In the present study, MTurk-based responses for a personality scale were found to be significantly less reliable than scores previously reported for a community sample. While score reliability was not affected by the length of the survey or the payment rates, the presence of an item asking respondents to affirm that they were attentive and honest was associated with more reliable responses. Best practices for MTurk-based research and continuing research needs are addressed.

1. Introduction

Psychological researchers have long been open-minded in adopting new technologies to aid in the process of conducting research, while staying committed to protecting the integrity of that process. This openness has been evident in the use of computer-banks requiring punch-card data entry in the 1950s (Stevens, 1951), the adoption of "mini-computers" to run experiments in the 1970s (Castellan, 1975), the administration of personality tests via computers in the 1980s (Butcher, 1987), and the use of online surveys for data collection purposes at the start of the 21st century (Birnbaum, 2000). In each case, psychologists have harnessed the power and promises of technology, while also critically evaluating possible weaknesses and ways that the technologies might change the research processes; it has been a history in which optimism has been balanced with caution.

Amazon's Mechanical Turk (MTurk) represents a new technological advancement that is already being adopted by psychological researchers because of the promise it holds. MTurk was introduced by Amazon in 2005 as a "marketplace for work that requires human intelligence" (www.mturk.com), pairing together "requestors" and "workers" for short-term tasks. Requestors make a payment to Amazon to post Human Intelligence Tasks (HITs), which are self-contained jobs that may be as short as 1 min; examples of HITs include identifying objects in pictures, transcribing unclear copies of receipts, or visiting a website to comment of the clarity of the information. The requestor then determines the number of workers who can be paid for the HIT, the required qualifications of the workers, and the "reward" or payment that will be given to a worker upon completion of the HIT. For example, a requestor might agree to pay 100 workers a $.10 reward each for the 1-min task of visiting two different sample websites and rating the esthetic appeal of each. The HIT is then posted to any registered workers who meet the required qualifications, and payment is transferred to the workers' MTurk accounts at the completion of the task.[1]

Shortly after its introduction, psychologists began to see MTurk as a means of gaining rapid access to a large and diverse sample of research participants. In what appears to be the first psychological research publication using MTurk workers as participants, Eriksson and Simpson (2010) found that both American and Indian women (relative to American and Indian men) expressed stronger negative emotional reactions to the prospect of losing money in a lottery, replicating a gender difference in financial risk-aversion in an online setting. Soon after, Alter, Oppenheimer, and Zemla (2010) used MTurk workers to study the tendency to overestimate one's ability to explain natural and mechanical processes. In this study, MTurk workers showed the same cognitive error that had previously been observed in lab settings.

The first appearance of MTurk data in the research literature was closely followed by the first peer-reviewed article examining the strengths and weaknesses of this source of data. Paolacci, Chandler, and Ipeirotis (2010) examined the demographic characteristics of 1000 MTurk workers. Their results showed that, at that time, nearly half were U.S. residents, but with a large proportion of

E-mail address: steve.rouse@pepperdine.edu

[1] For more information about the details of using MTurk, see Mason and Suri (2012) and Paolacci and Chandler (2014).

http://dx.doi.org/10.1016/j.chb.2014.11.004
0747-5632/© 2014 Elsevier Ltd. All rights reserved.

residents in India as well. Approximately two-thirds were women. Although the education level of the U.S. workers was higher than that of the general U.S. population, the reported salary was lower, and the majority of the workers completed the HIT for purposes of entertainment or to pass time. In a replication of classic decision-making research tasks, Paolacci and colleagues showed that rates of judgment errors did not differ significantly for MTurk workers in comparison to more conventional samples. Paolacci and colleagues concluded that MTurk offered numerous benefits, including the lack of potentially-biasing interactions with the experimenter, the access to a distinctly different population than traditional college samples, built-in anonymity, and availability of cross-national data. A potential concern about inattentiveness led to a suggestion that studies include attention-checks; their survey included one such check in the form of a question that asked "While watching the television, have you ever had a fatal heart attack?" (p. 416). They also noted that motivation, attrition rates, and attention could be influenced by providing fair rewards; they paid $.10 to respond to a survey that required fewer than 5 min to complete.

In what has arguably become the most authoritative endorsement of MTurk data, Buhrmester, Kwang, and Gosling (2011) argued that MTurk samples were more demographically diverse than college samples and that participation rates varied on the basis of reward rates and estimated task-completion times. Especially notable, Buhrmester and colleagues demonstrated that their MTurk data met acceptable psychometric quality criteria. In three different HITs, with rewards set at $.02, $.10, and $.50, responses were collected for six personality tests. Alpha coefficients ranged from .73 to .93; although the alphas did vary somewhat across the different payment levels, the mean alphas were judged to be very close to reliability estimates obtained for conventional samples (although the differences in reliability estimates were not tested for statistical significance). Similarly, 3-week test–retest reliability estimates ranged from .80 to .94, similar to those obtained for other samples. Buhrmester and colleagues concluded that MTurk data has great promise for conducting research in psychology, though they noted that "the process of validating MTurk for use by researchers has only just begun" (p. 5) Despite this call for continued critical evaluation of MTurk data, this article has arguably become a standard generalized endorsement of this method of data collection. For example, a PsycINFO[2] search of the 2013 publication year yielded 161 sources that cited Buhrmester et al. (2011). Many of these cited this article as a general source for the method (e.g., "...each word was rated by 40 Amazon Mechanical Turk workers (see Buhrmester et al., 2011)" (Adelman & Estes, 2013, p. 531)) without addressing questions about the quality of the data. Others used this source as a general statement to support the quality of MTurk data (e.g., "Mechanical Turk is an online survey service hosted by Amazon, where workers complete the tasks for nominal fees and provides similar results to other online and traditional recruitment methods (Buhrmester et al., 2011)" (Dailey, Brody, LeFebvre, & Crook, 2013, p. 1041)). Some authors specified the reward paid and calculated reliability for the scores obtained from the MTurk responses, while others neglected to report either or both pieces of information.

Although Buhrmester et al. (2011) have been cited more often than other researchers to support the use of MTurk data, a small number of additional researchers have also explored the psychometric reliability of MTurk data. For example, Johnson and Borden (2012) reported data for six different personality tests administered both to an MTurk sample and a lab sample; for some scales, the reliability estimates only varied by .01 across samples, while other scales had differences as large as .14 across samples, but these differences in reliability estimates were not tested for statistical significance. Also, Holden, Dennie, and Hicks (2013) examined the 3-week test–retest reliability of scores on a measure of the Five Factor Model of personality; correlations ranged from .79 to .91, and all paired-samples *t*-tests were nonsignificant. The results from both of these studies were broadly interpreted as support for the reliability of MTurk data.

Clearly, MTurk has quickly become an important development in the contemporary landscape of psychological research; unlike previous technological advances, however, the enthusiastic adoption of the method may have preceded the cautious evaluation of its strengths and weaknesses. Additional research is needed for several reasons. First, as Buhrmester et al. (2011) noted, the supportive data that they provided should not be considered the final word on validating MTurk as a means of gaining research data. Second, while the work of Buhrmester et al. (2011) and Johnson and Borden (2012) contrasted alpha reliability coefficients obtained for MTurk and conventional samples, the differences between these alpha coefficients were not tested for statistical significance but were subjectively judged on their degree of similarity. Third, the online supplemental material provided by Buhrmester et al. (2011) to accompany the published manuscript showed that the reliability coefficients obtained did fluctuate across the different reward levels, but no studies have systematically or statistically examined factors that might be associated with significantly higher or lower levels of reliability, such as the length of the HIT, the reward paid for the HIT, or attentiveness/accuracy checks for the HIT. The present study sought to contribute to the research literature by collecting data from different MTurk samples and varying different factors for the HITs in order to systematically evaluate differences in reliability coefficients.

2. Methods

2.1. Procedure

Eight different surveys were created within the MTurk website. All eight included demographic questions regarding age, sex, and race/ethnicity. In addition, all eight forms included a question to screen for inattentive responders and artificial intelligence systems ("bots") that are created to rapidly complete surveys: "What is the third word in this question: *How many stars are in the American Flag*?", with the correct response being "stars". In addition, all eight forms included the 10-item Openness to Experience (O) scale from the International Personality Item Pool (IPIP; Goldberg et al., 2006). This scale included five positively-keyed items and five negatively-keyed items, presented on a five-point Likert scale, with positively-keyed items scored from 1 (Strongly Disagree) to 5 (Strongly Agree), and negatively-keyed items scored from 5 (Strongly Disagree) to 1 (Strongly Agree); skipped items were recoded as "3". Goldberg (1999) reported a reliability estimate of .82 for a standardization sample of 501 adults from a community sample.

Although the eight forms were similar in many ways, they varied by three factors. First, half of the forms offered a $.10 reward for the completion of a 2–5 min survey (i.e., "Conventional reward form"), and the others offered a $.30 reward for the completion of a 2–5 min survey (i.e., "Generous reward form"). Second, half of the forms only included the 10-item O scale (i.e., "Short form"), and the other half included the 10-item O scale along with 10 filler items (not included in the analyses) from a 20-item version of the IPIP O Scale (i.e., "Long form"). Third, half of the forms concluded with a question asking participants to either affirm that they were

[2] It is important to note, however, that straightforward PsycINFO searches are likely to underestimate the number of researchers who have utilized MTurk samples; although 158 of the 161 articles that cited Buhrmester et al. (2011) in 2013 presented data collected from an MTurk sample, only 8 of them included the terms "MTurk" or "Mechanical Turk" or similar searchable terms in the abstract or title.