



# Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk<sup>☆</sup>



Timothy M. Daly<sup>a,\*</sup>, Rajan Natarajan<sup>b,1</sup>

<sup>a</sup> United Arab Emirates University, Abu Dhabi, United Arab Emirates

<sup>b</sup> Auburn University, Auburn, AL 36849, USA

## ARTICLE INFO

### Article history:

Received 1 March 2015

Received in revised form 1 April 2015

Accepted 1 April 2015

Available online 24 May 2015

### Keywords:

Longitudinal research

Crowdsourcing

Amazon Mechanical Turk

Data collection

Sampling

## ABSTRACT

Locating reliable sources of generalizable longitudinal data is an extremely important issue for business research. The aim of this paper was to empirically verify that crowdsourcing can be used to source longitudinal samples. Specifically, three studies assess reliability of the Amazon Mechanical Turk Marketplace (MTurk). All three studies demonstrate that MTurk is a reliable, inexpensive source for generalizable longitudinal data. Study 1 ( $n = 752$ ) examines the two-month re-response rate (study 1,  $n = 752$ ; 75%) of a US MTurk sample. Study 2 ( $n = 373$ ) investigates the four- and eight-month re-response rate (56 and 38%, respectively) of a US immigrant sample. Study 3 examines the thirteen-month re-response rate (47%). Each study demonstrates minimal non-response biases and longitudinal response consistency, in terms of both demographics and personality traits. This study also independently verifies the accuracy of self-report state of residence for 94% of the participants.

© 2015 Elsevier Inc. All rights reserved.

## 1. Swapping bricks for clicks: Crowdsourcing longitudinal data collection with Amazon Mechanical Turk

An opportunity for improving cross-sectional business research lies in the potential to further explore theories and issues with longitudinal research designs. Indeed, some theories and models inherently rely upon time-separated data from individuals. For example, brand loyalty and brand switching are vitally important to branding research, but are almost impossible to access without some type of temporally-separated design (e.g., Dawes, Meyer-Waarden, & Driesener, 2015). This type of research typically includes a true-panel design where the diagonal elements represent brand loyalty and the off-diagonal ones indicate extents of brand switching. Similarly, technology acceptance (e.g., Brown, Venkatesh, & Goyal, 2014; Venkatesh, Thong, & Xu, 2012), test-retest for scale development (see MacKenzie, Podsakoff, & Podsakoff, 2011), purchase intention-to-behavior relationships (e.g., Pavlou, Liang, & Xue, 2007), and pre- and post-communication campaign research (e.g., Johnston & Warkentin, 2010) are among other research topics that depend on multiple time-points.

Unfortunately, it is often very difficult to source a reliable and generalizable sample that can be dependably accessed across multiple time-

points. In fact, two dominant options for this type of sampling are currently available to the interested researcher: students and commercial research panels (this puts aside corporate samples, which are a more specific issue). The major benefits of recruiting students are low attrition rates (Bhattacharjee & Premkumar, 2004) and low costs, as students are generally paid in course credit or cheap prize draws. In contrast, the major benefits of commercial panels are increased generalizability and the ability to make specific requests regarding demographic, psychographic, or other segmentation bases.

Despite these benefits, student and commercial research panel samples have a number of significant disadvantages that make it necessary to explore other options. For student samples (non-probability convenience samples) these include low external validity and limited access for researchers outside of academia, or those at universities that discourage recruiting students for research (Mason & Suri, 2012). For commercial research panels, disadvantages include significant monetary costs coupled with little guarantee of usable re-response rates. For example, one major US-based research panel provider estimates a 50% re-response rate after two months but only 15% after 13 months for a nationally representative non-specific US sample. This is based upon estimated costs of \$5 per completed respondent at Time 1, increasing to \$7 and \$9 at each subsequent time period. Therefore, it is imperative to uncover new sources of longitudinal data, as neither of these two existing options can provide the caliber of solutions that high-level research requires.

The present research proposes and demonstrates that online crowdsourcing marketplaces have the same advantages as student samples and commercial research panels without their significant disadvantages. A crowdsourcing marketplace is essentially a digital

<sup>☆</sup> The authors thank Sarah Rasmi for her comments on an earlier version of this manuscript.

\* Corresponding author at: College of Business and Economics, United Arab Emirates University, Al Ain, UAE, P.O. Box 15551. Tel.: +971 3 713 5219.

E-mail addresses: tim.daly@uaeu.ac.ae (T.M. Daly), natarra@auburn.edu (R. Natarajan).

<sup>1</sup> Tel.: +1 334 844 2450.

labor market, wherein employers can contract anonymous workers to complete a task. The rationale behind this concept is that it is simpler and more accurate to have many individuals complete a large number of small tasks than to develop the complex algorithms and computer code that are required to automate the process. Typical tasks can include surveys (academic or professional), transcription of audio files, classification of digital information (such as receipts or websites), and tagging photos.

This study focuses on the Amazon Mechanical Turk Marketplace (MTurk) because MTurk is clearly the dominant platform on the market and has a strong brand history, which suggests that Amazon will support it for years to come. To evaluate the utility of MTurk, this study examines re-response rates (across time periods of two, four, eight and thirteen months), non-response biases, and the stability and consistency of objective (demographic) and subjective (Big-Five personality traits) self-report measures over time. Combined with a custom-built web application for bulk messaging within the MTurk system (available to academic researchers upon request from the first author), this research equips the reader with the tools to take full advantage of MTurk for longitudinal research projects.

## 2. Amazon's Mechanical Turk

### 2.1. The use of MTurk in academia

MTurk is rapidly becoming an influential source of non-student research samples (Goodman, Cryder, & Cheema, 2013; Rand, 2012). In order to use MTurk for sample recruitment, the researcher (in the role of “requestor”) must simply publish a job (referred to as a “HIT”) for employees (referred to as “workers”) and provide a payment rate with the survey link to the applicants. The Requestor has the option to specify a number of criteria to ensure a quality sample, including worker experience level, previous job acceptance rate, and residence country. MTurk attracts considerable academic interest given that it facilitates rapid recruitment at a much lower cost than commercial research panels (with the important related benefit of an in-built and flexible micro-payment system). This academic interest has covered a wide range of business topics and contexts including corporate social responsibility (Skarmas & Leonidou, 2013), consumer behavior (Xia & Kukar-Kinney, 2014), branding (Swimberghe, Astakhova, & Wooldridge, 2014), social media usage (Qiu, Lin, Ramsay, & Yang, 2012), decision-making (Fast, Sivanathan, Mayer, & Galinsky, 2012), consumer behavior (Collier & Barnes, 2015), scale development (Baldus, Voorhees, & Calantone, 2015), virtual work team relationship quality (O’Leary, Wilson, & Metiu, 2014), personality (Jones & Paulhus, 2011), and cognition (Paxton, Ungar, & Greene, 2012).

MTurk is a reliable source of participants for academic research (e.g., Mason & Suri, 2012; Sprouse, 2011). Research shows that U.S.-based MTurk workers report comparable scale reliabilities to US-based university students and general online panel provider samples (Buhrmester, Kwang, & Gosling, 2011; Steelman, Hammer, & Limayem, 2014). Other studies using U.S.-based MTurk workers were able to replicate theoretical models such as the conjunction fallacy and framing effects (Paolacci, Chandler, & Ipeirotis, 2010). Generally, MTurk Workers are comparable to diverse online panels (Steelman et al., 2014), making them more diverse than student samples (Buhrmester et al., 2011).

Intrinsic and extrinsic rewards motivate MTurk workers (Ipeirotis, 2010; Kaufmann, Schulze, & Veit, 2011). As a result, they are as attentive to research tasks as students and online panel samples (Paolacci et al., 2010). The right of requestors to withhold payment for poor quality work (which then has the added effect of negatively impacting the worker’s quality rating and thus prospects for future employment within these systems) is an important check-and-balance in relation to extrinsic motivation, reducing the likelihood of unreliable survey responses.

Despite a general upward trend in the use of MTurk to recruit research participants, only a handful of studies attempt to use the platform for any type of time-separated data collection. The few exceptions recollect data after three weeks or less (e.g., Holden, Dennie, & Hicks, 2013; Shapiro, Chandler, & Mueller, 2013). Overall, it is clear that researchers have avoided using MTurk for any sort of extended longitudinal research for several reasons. First, investing the time and money required to set up a research panel in MTurk is risky without any empirical data demonstrating the acceptability of re-response rates and non-response biases. Second, contacting participants individually via the MTurk system is extremely time consuming and cumbersome. Direct contact also directly violates the MTurk site use policy to request the email address of MTurk workers. This limitation is problematic because participants are unlikely to complete follow-up studies without notification (other than by chance). To address these issues, this study examines participant re-response rates over several time-points and potential non-response biases that dropouts introduce. This study also demonstrates that using a simple bulk messaging Python app (customized versions available to academic researchers upon request to the first author) facilitates re-contacting MTurk Workers.

### 2.2. Re-responses rates and non-response bias on MTurk

Assessing the expected level of participant re-response rate over time is fundamentally important when evaluating MTurk for longitudinal research. The few studies that report time-separated data on MTurk report reasonably high re-response rates over short time-periods. The reported re-response over a three-week period ranges from 60% (Buhrmester et al., 2011) to 69% (Holden et al., 2013). High (80% plus) re-response rates were reported over a one-week interval (Shapiro et al., 2013). However, most longitudinal business research requires more than a three-week time period. Therefore, this study examines the re-response rates at two-month (Study 1), four-month, eight-month (Study 2) and thirteen-month intervals (Study 3). This study also analyzes demographic information (gender, age, income, education) and Big-Five personality traits reported at Time 1 to identify any potential differences between re-responders and those who dropped out.

### 2.3. Participant temporal consistency

Verifying that MTurk participants provide consistent answers across timepoints is important, particularly on objective measures such as demographics. Many researchers have concerns that online participants are providing false or misleading information (Sprouse, 2011) because researchers have less control when using a purely online platform compared to in-person laboratory or classroom studies. Rand (2012) examines the consistency of MTurk responses for participants who coincidentally complete two of his posted MTurk HITS (Only 100 out of 3142 [3%] participants cross his two studies, no time period information is provided). Almost all participants in these studies report the same gender (96%) and age (93%) at two different data points. This consistency provides some evidence for reliable responding, given that self-reported demographics such as gender and birth year are expected to remain constant over time.

This study uses a similar method to investigate reliable responding by comparing relatively stable and enduring objective (i.e., demographics) and subjective (i.e., Big-Five personality traits) measures over time. Small changes in personality can occur, but this usually happens over long periods of time (e.g., Terracciano, McCrae, & Costa, 2010). Therefore, high test–retest reliability is a desirable quality for any personality scale (e.g., Milojev, Osborne, Greaves, Barlow, & Sibley, 2013). As the maximum time period of interest in this research is just over one year, using an established personality scale will enable a valid examination of the temporal consistency of subjective data.

In addition to examining response consistency, this study also considers non-self report data verification by comparing the geo-located

Download English Version:

<https://daneshyari.com/en/article/1017056>

Download Persian Version:

<https://daneshyari.com/article/1017056>

[Daneshyari.com](https://daneshyari.com)