Journal of Business Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Business Research

Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk[☆]

Timothy M. Daly [a,*], Rajan Nataraajan [b,1]

[a] United Arab Emirates University, Abu Dhabi, United Arab Emirates
[b] 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 xxxx

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.

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 timepoints.

Unfortunately, it is often very difficult to source a reliable and generalizable sample that can be dependably accessed across multiple timepoints. 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 (Bhattacherjee & 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 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

[☆] 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. Nataraajan).

[1] Tel.: +1 334 844 2450.

http://dx.doi.org/10.1016/j.jbusres.2015.05.001
0148-2963/© 2015 Elsevier Inc. All rights reserved.

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 (Skarmeas & 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-reponses 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 Internet Protocol (IP) address of participants with their self-reported location. Specifically, self-reported US state of residence (via zip code) is compared with the state data produced by a commercial IP geo-location service (MaxMind). This method is used to establish that the vast majority of participants accurately self-report their country of residence (97%; Rand, 2012).

2.4. The current research

This study examines the utility of MTurk as a source of longitudinal data across three studies. Study 1 assesses the practicality of using MTurk to collect short-term temporally separated data. This method can be used for a test–retest study, or when one wishes to separate certain scales from experimental manipulations. Specifically, this study quantifies the re-response rate at a relatively short two-month time interval using a sample of U.S.-based MTurk Workers. This study examines non-response bias using objective demographic indicators. Study 2 quantifies the re-response rate of MTurk participants at medium (four- and eight-month) time intervals using a sample of first- and second-generation US immigrants. This study examines non-response bias using both objective demographic and subjective Big-Five personality traits. Study 3 quantifies the re-response rate and associated non-response biases of US-based MTurk Workers over a longer thirteen-month time interval. This study also investigates response consistency using three separate indicators: objective response consistency (gender and birth year); subjective response consistency (Big-Five personality traits); and independent location data verification compared to self-reported country and state of residence.

3. Study one

3.1. Participants and procedure

Eight hundred members of the MTurk marketplace participated in Study 1 at time 1 (T1). All participants were recruited from a HIT posted to MTurk describing an academic survey about personal values, which was the subject of a separate study. This HIT specified a payment of USD $2.30 for an estimated 23-minute task. Although this payment is above the average hourly wage on MTurk, it is still significantly below the minimum wage in the USA. MTurk Workers regard a rate of 10c/min as borderline acceptable, without being unethically low (http://www.wearedynamo.org/). Participants were screened to ensure that they were located within the U.S., had previously completed at least 50 HITS, and had an MTurk job acceptance rate of greater than 96%.

Basic demographic information was collected at T1, including birth year, gender, residential zip code, education level, and annual household income. Two attention checks were also embedded in the personal values study (e.g., "It is important to show you are reading the items by selecting 'not like me at all' on the far left"). Forty-eight participants were excluded from further analysis as they failed one or both of the attention checks, resulting in a final sample of 752 participants. This sample was evenly distributed by gender (48% female) and had an average age of 35 years old (SD = 11.2). The majority (88%) had at a minimum some college education, and the median weekly household income range was $500 to $999 per week (see Table 1 for full demographics).

The 752 attentive participants were re-contacted after approximately two months via a custom built MTurk messaging web-application (see Fig. 1). Participants were invited to complete a follow-up HIT for a separate research project that was restricted to this sample and paid the same rate (10c/min). This HIT was made available for ten days, and a reminder message was sent via the app after three days to non-responders.

3.2. Attrition rate and response bias

Five hundred sixty-two (75%) of the 752 participants at T1 responded at Time 2 (T2). These participants were not significantly different from the overall sample at T1 with respect to gender ($\chi^2_{1,\ 562} = .86$, $p = .35$), education ($t_{561} = 0.57$, $p = .57$), or income ($t_{544} = -.80$, $p = .42$). However, participants at T2 ($M = 37$, $SD = 11.4$; $t_{559} = 2.94$, $p = .003$) were significantly older than the total sample at T1 ($M = 35$, $SD = 11.2$).

Table 1
Demographics of overall samples (T1) and respondent samples.

	Study 1		Study 2			Study 3	
	T1	T2	T1	T2	T3	T1	T2
Gender							
% Female	48	50	48	51	52	55	63
Age							
Mean	35	36	30	32	33	34	37
SD	11.2	11.4	10.5	11.4	12.3	11.6	12.5
Minimum	18.00	19.00	19.00	19.00	19.00	19.00	19.00
Maximum	73.00	73.00	65.00	65.00	65.00	71.00	71.00
Education Level							
1 Some high school	0.00	0.01	0.01	0.01	0.01	0.01	0.01
2 High school/GED	0.11	0.11	0.08	0.07	0.09	0.12	0.11
3 Trade tech school	0.01	0.01	0.01	0.01	0.01	0.02	0.02
4 Some college	0.29	0.27	0.39	0.34	0.33	0.32	0.28
5 Associate's degree	0.12	0.12	0.06	0.06	0.06	0.09	0.07
6 Bachelor's degree	0.36	0.37	0.34	0.36	0.35	0.33	0.36
7 Master's degree	0.09	0.10	0.09	0.13	0.12	0.09	0.13
8 Doctorate (e.g., PhD)	0.02	0.03	0.02	0.03	0.03	0.02	0.02
Mean	5.0	5.0	4.9	5.1	5.0	4.8	5.0
SD	1.5	1.5	1.5	1.5	1.5	1.6	1.6
Income per week							
1 < $150	0.05	0.07	0.05	0.05	0.07	0.08	0.05
2 $150 to $499	0.19	0.17	0.21	0.19	0.17	0.19	0.21
3 $500 to $999	0.28	0.29	0.26	0.28	0.29	0.34	0.36
4 $1000 to $1699	0.28	0.26	0.29	0.28	0.26	0.23	0.24
5 $1700 to $2499	0.12	0.12	0.14	0.12	0.12	0.10	0.09
6 $2500 plus	0.08	0.08	0.06	0.08	0.08	0.06	0.06
Mean	3.4	3.3	3.5	3.4	3.4	3.3	3.3
SD	1.2	1.2	1.3	1.3	1.3	1.3	1.2

3.3. Temporal response consistency

This study compares T1 and T2 responses to objective self-report data (i.e., gender and birth year) in order to assess participant consistency over time. Only a handful of participants reported a different gender ($n = 3$) and/or birth year ($n = 10$) across the time periods. Therefore, this study demonstrates a high re-response rate, low response bias and consistent responses over a relatively short time period. Study 2

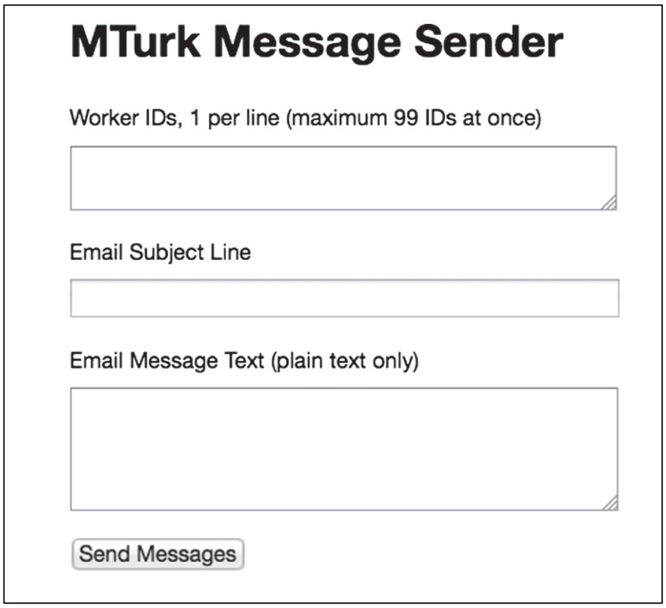
MTurk Message Sender
Worker IDs, 1 per line (maximum 99 IDs at once)
Email Subject Line
Email Message Text (plain text only)
Send Messages

Fig. 1. MTurk bulk messaging application user interface.

will investigate MTurk as a source of longitudinal data for a more specific targeted sample (US immigrants) over longer time-periods.

4. Study two

4.1. Participants and procedure

The ability to target and profile a specific sample make-up is one of the major benefits that commercial panel providers offer. Study 2 replicates this specific profiling in the MTurk environment. This study includes participants who were either first-generation (n = 136) or second-generation (n = 303) immigrants who had migrated to the U.S. from a non-English speaking country. This study sampled these participants from within a larger non-specific US-based sample (N = 2712). The HIT for this survey specified a payment of USD $1.80 for an estimated 20-minute task. Participants were screened to ensure that they were U.S.-based and had an MTurk job acceptance rate of greater than 96%.

Basic demographic information was collected at T1, including birth year, gender, residential state, education level, and annual household income. Participants were evenly distributed according to gender (48% female), with an average age of 30 years old (SD = 10.5). The vast majority (91%) of participants had at a minimum some college education, and the median weekly household income range was $500 to $999 per week (see Table 1 for full demographics). The 10-item Big-Five Inventory (BFI-10; Rammstedt, Goldberg, & Borg, 2010) was used to measure Big-Five personality traits on a seven-point scale (1 = "Strongly Disagree" and 7 = "Strongly Agree". Table 2 contains descriptive statistics for all variables that were used in this study.

Each participant was recontacted and invited to complete another HIT at four and eight months after initial recruitment. Both HITs paid the same as in Study 1 (10c/min) and were only accessible via direct invitation. Both HITs were made available for 10 days.

4.2. Attrition rate and non-response bias

Two hundred nine (56%) of the 373 participants responded at T2 and 141 (38%) responded at Time 3 (T3). Participants at T2 did not differ significantly from the overall sample for gender ($\chi^2_{1,\ 209} = .69$, $p = .41$) or income ($t_{184} = –.33$, $p = .74$). However, the participants at T2 were significantly older ($M = 32$, $SD = 11.4$; $t_{208} = 3.00$, $p = .003$) and more educated ($M = 5.1$, $SD = 1.5$; $t_{208} = 2.10$, $p = .037$) than the overall sample ($M = 30$, $SD = 10.5$; $M = 4.9$, $SD = 1.5$, respectively). Generally, participants at T2 did not differ from the overall sample with respect to personality traits. However, participants who re-responded were marginally more conscientious ($M = 5.4$, $SD = 1.2$; $t_{208} = 1.88$, $p = .06$) than the overall sample ($M = 5.2$, $SD = 1.2$).

The participants at T3 did not differ significantly from the overall sample for gender ($\chi^2_{1,\ 141} = .71$, $p = .41$), education ($t_{140} = .91$, $p = .36$) or income ($t_{132} = –.24$, $p = .81$). The participants at T3 ($M = 33$, $SD = 12.3$; $t_{139} = 3.08$, $p = .003$) were again significantly older than the total sample. As above, participants at T3 were marginally more conscientious ($M = 5.4$, $SD = 1.2$; $t_{140} = 1.74$, $p = .084$). Participants at T3 were also significantly more agreeable ($M = 5.3$, $SD = 1.2$; $t_{140} = 2.18$, $p = .031$) than the overall sample ($M = 5.4$, $SD = 1.2$). Therefore, this study demonstrates relatively high response rates, low response bias and consistent responses over two intermediate time periods with a specific demographic subsample. Study 3 will investigate MTurk as a source of longitudinal data at thirteen months. In addition Study 3 will examine response consistency of a subjective personality scale, as well as use non-self report data to externally verify the truthfulness of respondents' self-reported state of location.

5. Study three

5.1. Participants and procedure

Five hundred and twenty eight U.S.-based MTurk workers participated in Study 3 at T1. All participants were recruited from the same original HIT as Study 2, but without the immigration related country of origin screening. As in Study 2, basic demographic information was collected at T1, including birth year, gender, residential zip code, education level, and annual household income. The sample was evenly distributed by gender (55% female), with an average age of 34 years old (SD = 11.6). The vast majority (85%) of participants had at a minimum some college education, and the median weekly household income range was $500 to $999 (see Table 1 for full demographics). Big-Five personality traits were measured using the same 10-item Big-Five Inventory as in Study 2. Descriptive statistics are presented in Table 2.

Approximately thirteen months after initial recruitment each participant was invited via the MTurk messaging application to complete a follow-up HIT for a separate research project. This HIT asked the participants to again complete the demographic and personality questions. This HIT was specified to pay at the rate of 10c per predicted minute, and was only able to be viewed and completed by the initial 528 participants in order to avoid unrelated participants. The HIT was made available for 10 days.

5.2. Attrition rate and non-response bias

Two hundred forty nine (47%) of the 528 participants responded at T2. Participants at T2 did not differ significantly from the overall sample with respect to household income ($t_{229} = .41$, $p = .68$). Participants who were female (T1 = 55% female, T2 = 63% female, $\chi^2_{1,\ 249} = 5.86$, $p = .016$), older ($M = 37$, $SD = 12.\ 5$; $t_{248} = 3.91$, $p < .001$), and more educated ($M = 5.0$, $SD = 1.6$; $t_{246} = 2.12$, $p = .04$) were more likely to re-respond at T2 than the overall sample ($M = 34$, $SD = 11.6$; $M = 4.8$, $SD = 1.6$, respectively). As above, re-responders were more conscientious ($M = 5.4$, $SD = 1.2$; $t_{248} = 2.15$, $p = .033$) and agreeable ($M = 5.3$, $SD = 1.2$; $t_{248} = 1.98$, $p = .049$) compared to the overall sample ($M = 5.2$, $SD = 1.3$; $M = 5.1$, $SD = 1.2$, respectively).

Table 2
Descriptive statistics for the Big-Five Personality traits measured at T1 (Study 2 and 3), and test–retest correlations of the traits measured at T1 and again at T2 (Study 3).

	Study 2			Study 3		
	T1	T2	T3	T1	T2	r
	(N = 373)	(N = 209)	(N = 141)	(N = 528)	(N = 249)	(T1 & T2)
Extraversion	3.5	3.5	3.5	3.6	3.6	.79
	(1.5)	(1.6)	(1.6)	(1.6)	(1.6)	
Agreeableness	5.1	5.1	5.3	5.1	5.3	.72
	(1.2)	(1.2)	(1.2)	(1.2)	(1.2)	
Openness	5.2	5.1	5.2	5.2	5.0	.69
	(1.2)	(1.2)	(1.2)	(1.2)	(1.2)	
Conscientiousness	5.2	5.4	5.4	5.2	5.4	.73
	(1.2)	(1.2)	(1.2)	(1.3)	(1.2)	
Neuroticism	3.3	3.2	3.3	3.3	3.3	.74
	(1.4)	(1.3)	(1.3)	(1.5)	(1.5)	

5.3. Temporal response consistency

This study compares T1 and T2 responses to objective self-report data (i.e., gender and birth year) in order to assess participant consistency over time. Only a handful of participants reported a different gender ($n = 2$) or birth year ($n = 14$) across the time periods. This study also examines the correlations between the Big-Five personality traits at T1 and T2 in order to assess subjective self-report participant consistency. Results show high intercorrelations, ranging from $r = .69$ (openness to experience) to $r = .79$ (extraversion). The average correlation between the time points was $r = .73$ (see Table 2). In addition to examining response consistency, this study also considers non self-report data verification by comparing the geo-located IP address of participants with their self-reported location using MaxMind. Results show a very high (87%) match rate. Participants who were mismatched were either inaccessible using MaxMind (11%), located in a neighboring state (2%), or located more than one state away (3%). A second geolocation database (WhatIsMyIPAddress.com) was used to resolve 7.2% of the mismatched participants. Overall, a high level (94%) of consistency was found between self-reported state of residence and the geolocated state.

6. Discussion

This study demonstrates that MTurk can be useful for collecting reliable temporal data in longitudinal studies. Specifically, MTurk yields strong re-response rates across multiple time-points (ranging from 75% at two months to 47% at thirteen months) with minimal non-response biases. Importantly, as per available evidence, MTurk produced much larger re-response rates vis-à-vis commercial panel predictions (~15% at 13 months). Therefore, MTurk is far more accessible than financially prohibitive commercial options and more generalizable than student samples, opening up new avenues business research (see Steelman et al., 2014 for a full investigation comparing MTurk, commercial panels, and students).

Overall, the non-response biases are minimal. However, one relatively consistent bias emerges: Participants who responded at each secondary time-point are older than the total sample. This difference is significant, but the magnitude of the differences is small (ranging from 1 to 3 years). It is possible that this re-response bias is related to the second most common response bias: Participants who responded at each secondary time-point are more conscientious than the total sample. Research has shown that conscientiousness (i.e., dependable, organized, thorough, hard-working) increases as people age (Jackson et al., 2009). Therefore, it is likely that older participants are also more conscientious and therefore more likely to respond to later surveys. We find some support for this idea, as conscientiousness is positively correlated with age for the overall sample of both Study 2 ($r = .26$, $p < .001$) and Study 3 ($r = .25$, $p < .001$). The higher levels of conscientiousness for re-respondents have the added benefit of making it more likely that high quality data will be captured over time given their thorough and hardworking nature.

This research also demonstrates that MTurk can be used to collect longitudinal data from a specific demographic segment. In this study, we showed that a diverse sample of first- and second-generation US immigrants could be selected and accessed over time via the MTurk system. This diversity represents a significant advantage over both university students and commercial panels. It is noted that the US immigrant sample demonstrates lower re-response rates than the non-immigrant US sample did over a longer time-period. However, this response rate is still well above useable levels (38% after eight months). A possible explanation is the relative youth of the overall sample (compared to the general samples in Study 1 and 3). Future research should examine re-response bias in specific segments to better understand the viability of using MTurk for segmented longitudinal research designs.

This study also yields very promising results in terms of response consistency and reliability using both objective, subjective, and geo-location indicators. First, the vast majority of participants report the same gender and birth year across time periods. Second, personality has a large positive correlation ($r_{\text{average}} = .73$) across the two time periods in Study 3. Third, the self-reported state of residence can be externally verified for almost all Study 3 participants. Taken together, this study presents strong evidence that MTurk users are genuine responders who provide careful responses, further strengthening the case for collecting longitudinal data using MTurk.

6.1. Research implications

First and foremost, this research clearly established that MTurk can be used in longitudinal research with both short- and relatively long timeframes. This finding corroborates previous work identifying the benefits of crowdsourcing relative to student and commercial research panel samples (Steelman et al., 2014). MTurk offers several advantages over student samples. Specifically, MTurk enhances external validity by targeting all ages, genders, income levels, and educational backgrounds, as well as facilitating participant segmentation (e.g., a specific income bracket, educational cohort, or ethnocultural group). MTurk also offers a number of advantages over commercial research panels. Specifically, MTurk is associated with decreased costs (compared to a typical minimum of $5 per participant from commercial research panels) and increased response rates (e.g., 47% at thirteen months compared to an estimated 15% from a commercial research panel). As a result, researchers should feel more confident investing their time and money into MTurk.

It should be noted that interested researchers could use MTurk to establish their own research panel. The process would require collecting a short and basic demographic and/or psychographic profile across a large pool of participants, with the objective of re-sampling from this pool. Researchers could then drill-down into the pool to extract specific segments of interests. This would essentially replicate the offerings of commercial research panels, without the high costs associated with longitudinal designs and sample specificity. Interested researchers can contact the corresponding author for more detailed information on how to execute this.

The results of these studies are particularly salient to business research. For example, this study demonstrates excellent viability for test–retest or short-term construct separation (75% at two months). Given that test–retest is an accepted prerequisite for valid scale development (MacKenzie et al., 2011) this new data source provides an opportunity to validate new and existing measures without relying upon student samples or the high costs associated with consumer research panels. In addition, high re-response rates at two, four, eight, and thirteen months that are demonstrated by this study can benefit researchers who are interested in topics that suited to a longitudinal design (e.g., technology adoption and acceptance). Therefore, MTurk is a strongly recommended tool for business research that calls for reliable and valid longitudinal research that is also relatively inexpensive and uses non-student samples.

6.2. Limitations and future directions

This investigation is limited to MTurk because MTurk is currently the most popular crowdsourcing platform in the world. Future research should study the viability of alternative US-based platforms (e.g., Crowdflower.com) as well as those based in other countries (e.g., CrowdGuru.com in Germany). These comparative assessments would facilitate informed decision-making and potentially trigger the emergence of even better platforms. Re-response data is collected at two, four, eight, and thirteen months. Future research should involve even longer durations, particularly in the temporal tracing of some branding aspects such as brand image, brand power, brand love, brand equity, and customer equity. Longer time-intervals could be used to reproduce theoretical models (Paolacci et al., 2010; Steelman et al., 2014) and shed more light on the stability of such reproductions as measured at different points in time.

This investigation is limited to U.S.-based users. The majority of business research can be conducted using this type of sample. However, it may be useful to consider other nationalities and also conduct cross-cultural comparative research. Steelman et al. (2014) report that non-U.S. MTurk users produce different model results to their U.S.-based counterparts and caution the use of these participants pending further investigation. Therefore, it is imperative that future research establishes which nations are both reasonably represented on MTurk and have acceptable re-response rates in order to facilitate this cross-cultural research.

Future research could also explore the impact of technology literacy on crowdsourcing users. After all, being tech-savvy is rapidly becoming a 21st century requirement for people in general (Nataraajan, 2014), but this may certainly vary both across and within cultures. It is a logical assumption that crowdsourcing users are more technology literate and accepting of new technologies than the average person, though the difference is less obvious when compared to students. Investigating this issue is an important next step in verifying the use of crowdsourcing for academic research.

Aspects of payment to participants and the resulting impact on them also pose interesting avenues to explore for researchers. For instance, would escalating the payments to Workers increase the likelihood of responding over time? An interesting investigation would be to state upfront in initial recruitment that payment for subsequent studies will be increased. This may have the benefit of encouraging users to track the requestor and look for opportunities to continue participating.

6.3. Practical implications

This research indicates two potential guidelines for best practices in the use of MTurk for longitudinal data collection. First, payment levels have to be made more attractive than current prevailing levels in order to ensure credible and valid data (although some concern regarding potential self-selection bias through attraction of professional survey participants may exist). Second, researchers should also seek to treat participants with respect. Prompt payment, answering questions and making surveys user-friendly are just some of the simply ways that researchers can help their Workers. Unmotivated Workers are more likely to provide careless, random or misleading responses (Huang, Curran, Keeney, Poposki, & DeShon, 2012) or may drop out.

Longitudinal research cannot be successfully executed without notifying participants about upcoming studies. This study uses a simple Python-scripted web application at a bulk level. It is not possible to offer a publicly available version of this application since this type of script requires both an individual's Amazon Web Services Access Key ID and Secret Access Key. However, interested academic researchers are encouraged to contact the corresponding author for a personalized messaging application. It is also possible to complete this manually through the MTurk system, though this is tedious and time-consuming.

6.4. Conclusion

Longitudinal studies (true-panel designs in particular) in business research are strongly needed. Indeed, a deeper understanding of branding contexts, consumption over time, investment behaviors over time, organizational behaviors over time, and assessing temporal efficacy of programs would benefit from longitudinal research. Unfortunately, the commonly used captive student samples and commercial research panels (now mostly online) have major disadvantages; external validity concerns with the former and cost concerns with the latter. In contrast, MTurk can be used to access longitudinal data that is reliable, valid, consistent, and inexpensive without relying on student samples. The studies here provide preliminary but meaningful testimony to this generalization.

References

Baldus, B. J., Voorhees, C., & Calantone, R. (2015, May). Online brand community engagement: Scale development and validation. *Journal of Business Research*, *65*(5), 978–985.

Bhattacherjee, A., & Premkumar, G. (2004, June). Understanding changes in belief and attitude toward information technology usage: A theoretical model and longitudinal test. *MIS Quarterly*, *28*(2), 229–254.

Brown, S. A., Venkatesh, V., & Goyal, S. (2014, September). Expectation confirmation in information systems research: A test of six competing models. *MIS Quarterly*, *38*(3), 729–756.

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

Collier, J. E., & Barnes, D. C. (2015, May). Self-service delight: Exploring the hedonic aspects of self-service. *Journal of Business Research*, *68*(5), 986–993.

Dawes, J., Meyer-Waarden, L., & Driesener, C. (2015). Has brand loyalty declined? A longitudinal analysis of repeat purchase behavior in the UK and the USA. *Journal of Business Research*, *68*(2), 425–432.

Fast, N. J., Sivanathan, N., Mayer, N. D., & Galinsky, A. D. (2012). Power and overconfident decision-making. *Organizational Behavior and Human Decision Processes*, *117*(2), 249–260.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013, July). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224.

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5–120 on Amazon's Mechanical Turk. *Computers in Human Behavior*, *29*(4), 1749–1754.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. *CeDER-10–01 working paper*. New York University.

Jackson, J. J., Bogg, T., Walton, K. E., Wood, D., Harms, P. D., Lodi-Smith, J., et al. (2009). Not all conscientiousness scales change alike: A multimethod, multisample study of age differences in the facets of conscientiousness. *Journal of Personality and Social Psychology*, *96*(2), 446–459.

Johnston, A. C., & Warkentin, M. (2010, September). Fear appeals and information security behaviors: An empirical study. *MIS Quarterly*, *34*(3), 549–566.

Jones, D. N., & Paulhus, D. L. (2011). The role of impulsivity in the dark triad of personality. *Personality and Individual Differences*, *51*(5), 679–682.

Kaufmann, N., Schulze, T., & Veit, D. (2011). More than fun and money. Worker motivation in crowdsourcing: A study on Mechanical Turk. *Proceedings of the 17th Americas Conference on Information Systems, Detroit, MI, August 4–7*.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011, June). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, *35*(2), 293–334.

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

Milojev, P., Osborne, D., Greaves, L. M., Barlow, F. K., & Sibley, C. G. (2013). The Mini-IPIP6: Tiny yet highly stable markers of big six personality. *Journal of Research in Personality*, *47*(6), 936–944.

Nataraajan, R. (2014). Likely research trends in the 21st century: A *ceteris-paribus* view from the 2014 marketing ivory tower. *Invited Talk given in June 2014 at the University of Braunschweig, Germany*.

O'Leary, M. B., Wilson, J. M., & Metiu, A. (2014, December). Beyond being there: The symbolic role of communication and identification in perceptions of proximity to geographically dispersed colleagues. *MIS Quarterly*, *38*(4), 1219–1243.

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

Pavlou, P. A., Liang, H., & Xue, Y. (2007, March). Understanding and mitigating uncertainty in online exchange relationships: A principal-agent perspective. *MIS Quarterly*, *31*(1), 105–136.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163–177.

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*(6), 710–718.

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of big-five factor markers for persons with different levels of education. *Journal of Research in Personality*, *44*(1), 53–61.

Rand, D. G. (2012, April). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013, April). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220.

Skarmeas, D., & Leonidou, C. N. (2013). When consumers doubt, watch out! The role of CSR skepticism. *Journal of Business Research*, *66*(10), 1831–1838.

Sprouse, J. (2011, March). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Steelman, Z. R., Hammer, B. I., & Limayem, M. (2014, June). Data collection in the digital age: Innovative alternatives to student samples. *MIS Quarterly*, *38*(2), 355–378.

Swimberghe, K. R., Astakhova, M., & Wooldridge, B. R. (2014). A new dualistic approach to brand passion: Harmonious and obsessive. *Journal of Business Research*, *67*(12), 2657–2665.

Terracciano, A., McCrae, R. R., & Costa, P. T., Jr. (2010). Intra-individual change in personality stability and age. *Journal of Research in Personality*, *44*(1), 31–37.

Venkatesh, V., Thong, J. Y., & Xu, X. (2012, March). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178.

Xia, L., & Kukar-Kinney, M. (2014). For our valued customers only: Examining consumer responses to preferential treatment practices. *Journal of Business Research*, *67*(11), 2368–2375.

Please cite this article as: Daly, T.M., & Nataraajan, R., Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk, *Journal of Business Research* (2015), http://dx.doi.org/10.1016/j.jbusres.2015.05.001