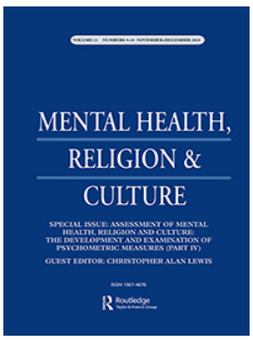

Routledge
Taylor & Francis Group

Mental Health, Religion & Culture

ISSN: 1367-4676 (Print) 1469-9737 (Online) Journal homepage: https://www.tandfonline.com/loi/cmhr20

Who is Mturk? Personal characteristics and sample consistency of these online workers

Martin J. Burnham, Yen K. Le & Ralph L. Piedmont

To cite this article: Martin J. Burnham, Yen K. Le & Ralph L. Piedmont (2018) Who is Mturk? Personal characteristics and sample consistency of these online workers, Mental Health, Religion & Culture, 21:9-10, 934-944, DOI: 10.1080/13674676.2018.1486394

To link to this article: https://doi.org/10.1080/13674676.2018.1486394

MENTAL HEALTH, RELIGION & CULTURE
2018, VOL. 21, NOS. 9–10, 934–944
https://doi.org/10.1080/13674676.2018.1486394

Routledge
Taylor & Francis Group

Who is Mturk? Personal characteristics and sample consistency of these online workers

Martin J. Burnham*, Yen K. Le* and Ralph L. Piedmont *

Department of Pastoral Counseling, Loyola University Maryland, Columbia, MD, USA

ABSTRACT

Amazon Mechanical Turk (MTurk) has increasingly attracted the attention of researchers as a convenient method to gather human subjects due to its efficiency, low cost, and ease of use in securing data samples. The current study provides a meta-analytic review across three separate MTurk samples. Participants were 1707 US citizens (774 males and 933 females, aged 18–77 years) recruited from Amazon's MTurk system. Results indicated that across three MTurk samples, demographic characteristics of workers closely approximated the general US population on gender and race but differed on religious affiliation given the very high number of Atheists and Agnostics (38.3%). Across the three samples, significant mean level differences were observed for 14 of the 15 study scales; however, effect sizes were small (eta^2 ranged from .01 to .04). Results suggested that individuals working in the fields of personality and spirituality need to be aware that using an MTurk sample may introduce sample bias in their data.

ARTICLE HISTORY
Received 31 May 2018
Accepted 5 June 2018

KEYWORDS
Amazon Mechanical Turk (MTurk); data collection; spirituality; personality

Amazon Mechanical Turk (MTurk) is an online data collection service run by Amazon.com which recruits research participants for small financial payments. MTurk has increasingly attracted the attention of social science researchers as an economical means to gather human research subjects due to its efficiency, low cost, and ease of use in securing participants. There are two types of participants in MTurk: "requesters" (task creators) and "workers" (paid task participants). Requesters can create and post any task such as surveys, experiments, and writing samples on the website. Workers can conduct tasks (known as Human Intelligence Tasks – HITs) requesters post in MTurk and are paid upon successful completion of each HIT. However, MTurk represents a contrived sample, in that individuals self-select to participate in this workforce and Amazon applies its own criteria for evaluating and retaining workers (e.g., Requester approval ratings). While MTurk provides requesters with access to a very large, readily available participant pool, some important questions remain regarding MTurk's sampling adequacy: What kind of participants do researchers secure when they do this sampling? Is their sample representative? How consistently will MTurk generate these representative samples? The current study attempts to address these questions by examining sample characteristics across three, large samples recruited over a six-month interval in an

CONTACT Ralph L. Piedmont mrmagic328@comcast.net
*All authors contributed equally to this project.

effort to examine which worker characteristics remain stable and which vary across studies. For social science researchers working with trait-type constructs, it would be important to know if any systematic demographic patterns or personality profiles that characterise these workers exist.

MTurk and its workers

According to Bohannon (2016), Google Scholar listed just 61 studies utilising MTurk in 2011; that number increased to 1120 studies in 2015. MTurk contains the major elements for conducting research such as an integrated participant compensation system, a large participant pool, and a streamlined process of study design, participant recruitment, and data collection (Buhrmester, Kwang, & Gosling, 2011). Traditional data collection, relying on undergraduate participant pools, has often confronted researchers with the problem of securing data in a timely and economical manner. MTurk crowdsourcing initiatives afford researchers alternatives for collecting data. MTurk participation is affected by compensation rate and HIT length. The median hourly wage for HITs performed on MTurk is $1.38 (Paolacci, Chandler, & Ipeirotis, 2010). For a short survey taken in about 2–4 minutes, MTurk respondents were paid $.25 (Berinsky, Huber, Lenz, & Alvarez, 2012). These low costs per participant are much less expensive in comparison to the cost for undergraduate samples ($5–10), for non-student campus samples (about $30), and for temporary agency subjects (about $15–20) (Kam, Wilking, & Zechmeister, 2007). Additionally, MTurk workers were not willing to complete long HITs for no or small compensation (Buhrmester et al., 2011). Compensation level and survey length do influence data collection speed but not data quality (Buhrmester et al., 2011).

Previous research has demonstrated that MTurk provides subject pools that move beyond the traditional undergraduate sample to offer wider demographics such as national and international samples of differing age, socioeconomic status, gender, and educational levels (Buhrmester et al., 2011). For example, in a sample of approximately 1500 MTurkers, Michel, O'Neill, Hartman, and Lorys (2017) reported the following sample demographics: the average age of MTurkers was 35.5 years (SD = 11.0); 58% of workers were female; 32.4% were single, 43.5% married, 15.5% living with a partner/significant other, and 8.5% separated/divorced/widowed; 75.5% were White/non-Hispanic, 8.9% African American, 6.2% Hispanic/Latino, 6.2% Asian, .8% Native American, and 2.4% other; the sample was educationally diverse with 26.1% having no degree, 16.0% having an Associate's degree, 40% a Bachelor's degree, and 17.9% an advanced degree. MTurkers have recently shifted to foreign countries, allowing researchers to include international subject pools. Of the subjects from 66 countries, the US accounted for the majority of MTurkers (46.80%), followed by India (34%), and other countries (19.20%) (Ipeirotis, 2012). Requesters can select the countries from which workers can be recruited.

There were both commonalities and differences reported in demographics between MTurkers and community samples, traditional social science research samples, internet samples, non-student adult samples, and the US population. Demographics of the MTurk sample, such as gender, race, and education, were more representative of the US population than traditional subject pools (Paolacci et al., 2010). Commonalities were shared between community samples and MTurkers on age and gender (Goodman, Cryder, & Cheema, 2012). When compared to traditional social science samples, MTurkers

shared similar demographics on race (Berinsky et al., 2012; Buhrmester et al., 2011) and gender (Buhrmester et al., 2011). In comparison to internet samples, MTurk samples were found to be similar in race (Berinsky et al., 2012) and gender (Buhrmester et al., 2011). When compared to student samples, MTurk samples had slightly more female participants (60.1%) than males (Berinsky et al., 2012; Ipeirotis, 2012). For religious affiliation, MTurk participants were more likely to report no religious affiliation (41.8%) when compared to traditional social science (26.9%) and online participants (13.1%; in Berinsky et al., 2012; see also Lewis, Djupe, Mockabee, & Su-Ya Wu, 2015).

Some researchers have noted demographic differences between MTurkers and the US population on age (Bohannon, 2016; Horton & Chilton, 2010; Mason & Suri, 2011) and gender (Mason & Suri, 2011); between MTurkers and internet samples on age (Buhrmester et al., 2011); and between MTurkers and a US college sample on age and gender (Goodman et al., 2012). While MTurkers (54% between 21 and 35 years) were found younger than the general US population (22% in this age range), MTurkers (mean years 32.3) were older than an average student sample (mean years 20.3), but younger than non-student adult samples (mean years 45.5; in Ipeirotis, 2012).

In terms of personality research, the MTurk sample was found equivalent to traditional samples from native-English speaking countries on the Big-Five personality dimensions. However, for non-English speaking participants, the MTurk sample was found nonequivalent to the traditional samples (Feitosa, Joseph, & Newman, 2015). The non-English MTurk sample scored significantly lower than the community sample on Extraversion, Emotional Stability, and Openness to Experience (Goodman et al., 2012). Furthermore, the higher percentage of Atheists and Agnostics may create restrictions in range for scores obtained from measures of spirituality and religiousness (Lewis et al., 2015).

What the above research indicates is that there can be important demographic and personological differences between MTurk workers and subjects obtained from more traditional recruitment strategies. As such, requesters may need to be more proactive and explicit in recruiting MTurk workers, in order to ensure that samples are representative of their intended populations and to avoid any restriction of range in scores obtained from measured personality-type variables.

The current study

Much of the research to date has reported on the findings from single data sets to various normative groups in the US. What seems to be lacking in the literature is any systematic analysis of MTurk samples across multiple samples collected at different points in time (however, see Lewis et al., 2015). Such a meta-analytic approach would allow for a better understanding of any systematic patterns of change and stability in the demographic characteristics and mean level response patterns on measured variables. The current study provides such a multi-study perspective by examining the demographic characteristics and mean level scores across three separate MTurk samples. Two of these samples were collected within a week of one another; the third sample was collected six months later. This collection pattern allowed us to determine the extent of sampling consistency and possible seasonal effects that may be operating in the data.

Our particular interest in this study was to determine the utility of MTurk samples with studies that included measures of personality and spirituality/religiousness. Do MTurk samples provide the type of representativeness conducive for research in this area (i.e., determining if there are any particular profiles on these dimensions that define MTurkers)? By using scales of personality, spirituality, and some psychosocial variables, the aim of the current study was threefold: (1) to validate whether MTurk samples were consistently demographically representative of the general US population; (2) to determine the value of MTurk crowdsourcing as a tool for data collection in spiritual/religious and personality inquiry, specifically to examine if there were any systematic personality profiles of these workers and whether the constructs of religiousness and spirituality were relevant to this population; and (3) to examine if there were any systematic patterns of consistency and change on the measured variables across multiple MTurk samples collected at different points in time.

Method

Participants

A total of 1707 participants, 774 males and 933 females, were recruited from Amazon's MTurk system for three independent studies, under the restriction that they were US residents, had English as their first language, and had at least a 95% task approval rating for their previous HITs. The age range for the total sample was from 18 to 77 years ($M = 35.74$, $SD = 12.03$). More specific participant characteristics for each study can be found in Table 1.

Measures

International Personality Item Pool 50 (IPIP-50)

Developed by Goldberg (1992), this instrument assessed the five major factors of personality with 50 items. The Big-five personality dimensions are Emotional Stability (ES), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). Each factor consists of 6 facet scales, each defined by 10 items. The IPIP items were administered with a five-point, Likert-type scale ranging from 1 ("very inaccurate") to 5 ("very accurate") as in the original instrument (Goldberg, 1992). Several items on this scale are reverse coded to account for acquiescence.

Assessment of Spirituality and Religious Sentiments (ASPIRES)

Developed by Piedmont (2010), this 35-item scale measures two major numinous dimensions: Spiritual Transcendence (STS) and Religious Sentiments Scales (RSS). The 23-item Spiritual Transcendence Scale includes three subscales: Universality (a measure of belief in the purpose of life and unity), Prayer Fulfilment (a feeling of joy and contentment resulting from prayer and/or meditation), and Connectedness (how much one feels responsible for and connected to others). The 12-item Religious Sentiments Scale includes two subscales: Religious Involvement (shows how actively involved a person is in religious rituals and activities) and Religious Crisis (measures how a person may experience problems, difficulties, or conflicts with the God of their understanding). The STS items are

Table 1. Sample demographics.

Demographics	Overall group $N = 1707$ %	Group 1 $N = 494$ %	Group 2 $N = 509$ %	Group 3 $N = 704$ %	χ^2
Gender					
Males	45.3	47.6	48.3	41.6	6.76*
Females	54.7	52.4	51.7	58.4	
Race					
Arabic	.5	.8	.4	.3	23.11*
Asian	6.3	8.1	4.3	6.4	
Black	7.9	9.5	6.1	8.0	
Caucasian	76.8	72.3	81.7	76.4	
Hispanic	5.3	5.3	4.7	5.8	
Mixed	2.6	3.8	2.0	2.0	
Other	.8	.2	.8	1.1	
Religious affiliation					
Catholic	17.1	20.0	16.3	15.6	76.64***
Unitarian	.5	.6	.2	.6	
Other Christian	15.2	13.2	15.9	16.1	
Buddhist	2.0	2.6	1.4	2.0	
Lutheran	2.0	3.2	2.2	1.0	
Baptist	9.4	11.3	9.0	8.2	
Jewish	1.7	2.0	1.8	1.4	
Methodist	3.5	2.6	5.3	.6	
Presbyterian	1.8	2.2	1.6	2.7	
Muslim	.9	1.6	.4	1.6	
Other faith tradition	5.6	9.1	6.3	.9	
Episcopal	.6	.2	.6	2.6	
Mormon	.9	.8	.8	1.0	
Hindu	.6	.8	.6	.9	
Atheist/Agnostic	38.3	29.4	37.7	44.9	

*$p < .05$, ***$p < .001$, two-tailed.

scored on a five-point Likert-type scale ranging from "strongly disagree" (1) to "strongly agree" (5), whereas the RSS items were scored on a seven-point Likert-type scale ranging from "never" (1) to "several times" (7). Piedmont (2010) provided evidence of reliability and validity for scores from this scale.

Affect Balance Scale (ABS)

Developed by Bradburn (1969), this scale measures the psychological well-being of respondents, assessing the occurrence of positive and negative events over the previous week. This scale contained 10 items composing three subscales: five items for positive affect (ABS Positive), five items for negative affect (ABS Negative), and affect balance (ABS Balance), which is computed by subtracting the negative item scores from the positive item scores. Scores on the ABS have demonstrated covariance with global happiness (Lowenthal, Thurner, & Chiriboga, 1975).

Purpose In Life Test (PIL)

Developed by Crumbaugh (1968), this 20-item scale measures a person's search for meaning and purpose in their lives as defined by Frankl (1985). Responses are given on a seven-point Likert-type scale, the poles of which vary according to the question. Scores can range from 20 to 140. Guttmann (1996) provided an adequate review of the research literature on this scale.

Procedure

This study was approved by Loyola University Maryland's Institutional Review Board. All four measures used in the study were digitised and loaded into a Qualtrix Survey. The survey was then uploaded onto Amazon's MTurk platform, an online platform for recruiting research participants. The first two samples of participants were gathered over two days (5 and 7 January 2016); both samples were completed in eight hours. The first sample resulted in 494 valid responses; the second sample yielded 509 valid responses. A third sample, collected in June 2016, was also completed in eight hours and yielded 704 valid responses. Workers who did not answer all questions or failed to answer the validity questions in the appropriate direction were discarded from the data set. Participants were each paid $1.00 US for their participation in the survey and were required to answer all questions in the survey to receive payment.

Results

The demographic characteristics of the MTurk sample are summarised in Table 1. The race of MTurkers was similar to the general US population. This MTurk sample consisted of 76.8% Caucasians (compared to 77.4% of US population, US Bureau of Labor Statistics, 2014). Furthermore, 7.9% were Black, 6.3% Asian, and 2.6% Mixed (compared to 13.2%, 5.4%, 2.5%, respectively, of the US population). However, in terms of religious affiliations, about 38.3% of MTurkers reported that they were Atheist or Agnostic, a percentage much larger than that found in the US population (4% in Lugo et al., 2008; 7.1% in Pew Research Center, 2014).

Pearson Chi-Square analyses showed a significant difference among the three MTurk groups for gender, χ^2 (2, $N = 1707$) = 6.76, $p < .05$; for race, χ^2 (12, $N = 1707$) = 23.11, $p < .05$; and for religious affiliation, χ^2 (28, $N = 1707$) = 76.64, $p < .001$. Residual analysis (i.e., the observed number of cases in each cell compared to their respective expected values) indicated a significant difference in religious affiliations for Lutherans in the second group (Std. Residual = 2.0) and in the third group (Std. Residual = −1.9); for Atheists/Agnostics in the second group (Std. Residual = −3.2) and in the third group (Std. Residual = 2.9); for methodists in the first group (Std. Residual = 2.2); and for Other Faith Tradition in the second group (Std. Residual = 3.3) and in the third group (Std. Residual = −3.4). These results indicated a significant difference between the observed and the expected frequencies of those aforementioned groups. A positive standardised residual indicated that Lutherans in the second group, Atheists/Agnostics in the third group, Methodists in the first group, and other Faith Tradition people in the second group were overrepresented in the sample. In contrast, a negative standardised residual suggested that Lutherans in the third group, Atheists/Agnostics in the second group, and Other Faith Tradition in the third group were underrepresented in those samples.

Worker IDs were evaluated to determine the extent of participation of workers across the three surveys. In comparing the first two samples, obtained within one week of each other, only nine individuals (approximately 1%) completed both surveys. Comparing the first and last surveys, completed approximately six months apart, 27 individuals (approximately 2%) participated in both. In comparing the second and last surveys, 30 workers (approximately 2.5%) were found to do both. Only three individuals participated

in all three surveys (less than 1%). Thus, the patterns of scores observed here can be considered qualities of the entire MTurk worker pool and not the result of a selection bias (i.e., results reflect only those interested in this type of research study).

Descriptive statistics and mean performance scores for study scales are listed in Table 2. Findings from a one-way Multivariate Analysis of Variance (MANOVA) indicated that across the different studies, there were significant study group differences on the measured constructs, Wilks Lambda = .95, multivariate $F(28, 3382) = 3.44$, $p < .001$.

Univariate one-way ANOVAs were performed for all the study variables and the results indicated that across the three studies there were significant differences for 14 of the 15 study scales; only the Positive Affect Scale showed no significant difference across three groups, $F(2, 1705) = .80$, $p > .05$. A Least Significant Difference (LSD) *post hoc* test was run to examine which samples exhibited significant mean differences on the measured variables and the results are presented in Table 2. The results of the LSD test indicated variability in the data among the three groups, although no identifiable pattern was noticed. However, the effect sizes for these differences range from an eta^2 of .01 to an eta^2 of .04; all values indicate small effects.

While mean scores varied across samples, a more fundamental issue concerned whether the interrelationships among the measured variables also varied across samples. Differing patterns of covariance would indicate that participants in the different samples understood the variables in different ways. This would create two problems. First, the lack of consistency in covariance would make aggregation of data across samples problematic. Second, it would undermine one's ability to replicate findings. A Box's Test of the Equality of Covariance Matrices was conducted to determine

Table 2. Results of mean performance scores on the study scales for the three groups.

	Overall group ($N = 1707$)		Group 1 ($n = 494$)	Group 2 ($n = 509$)	Group 3 ($n = 704$)		
Scales	*M*	*SD*	*M*	*M*	*M*	*F*	eta^2
Age	35.74	12.03	36.46	35.30	35.52	1.36	.00
Positive Affect	3.48	1.48	3.42	3.47	3.53	.80	.00
Negative Affect	2.51	1.59	2.29^{a}	2.73^{b}	2.52^{c}	9.75***	.04
Affect Balance	.97	2.37	1.14^{a}	.74^{b}	1.01^{ab}	3.71*	.01
Purpose in life test IPIP-50[a]	95.48	21.08	95.49^{ab}	93.21^{a}	97.07^{b}	4.87**	.02
Emotional Stability	52.31	13.59	53.92^{a}	50.93^{b}	52.11^{bc}	6.24**	.02
Extraversion	43.69	13.07	42.51^{a}	43.46^{ab}	44.70^{b}	4.24*	.01
Openness	54.53	12.83	52.69^{a}	55.03^{b}	55.52^{bc}	7.75***	.03
Agreeableness	47.61	12.63	46.44^{ab}	47.27^{b}	48.69^{bc}	4.97**	.02
Conscientiousness ASPIRES[b]	51.57	12.39	52.80^{a}	49.82^{b}	51.91^{ac}	7.75***	.03
Prayer Fulfilment	46.14	11.37	44.99^{a}	46.30^{ab}	46.85^{b}	4.06*	.01
Universality	48.69	9.74	47.59^{a}	48.98^{b}	49.29^{bc}	4.81**	.02
Connectedness	48.94	10.22	47.46^{a}	48.96^{b}	50.00^{bc}	9.18***	.03
STS Total Score	47.09	10.57	45.67^{a}	47.28^{b}	47.99^{bc}	7.25***	.02
Religious Involvement	41.81	11.32	40.48^{a}	42.71^{b}	42.14^{bc}	5.38**	.02
Religious Crisis	54.26	12.40	53.37^{a}	55.30^{b}	54.18^{ab}	3.05*	.02

Note: $N = 1707$, *$p < .05$, **$p < .01$, ***$p < .001$, two-tailed. Different superscripts denote significant differences among the three groups. STS Total Score = Spiritual Transcendence Total Scores.

[a]Scores were reported as T-scores with a mean = 50, and SD = 10 for the IPIP-50 (based on Roberts, Jangha, Piedmont, Sherman, & Williams, 2015).

[b]Scores were reported as T-scores with a mean = 50, and SD = 10 for the ASPIRES based on normative data (Piedmont, 2010).

whether the interrelationships among the measured variables fluctuated across samples and the results indicated no differences in the covariance matrices, Box's $M = 176.75$, multivariate $F(182, 6{,}560{,}076.61) = .961$, $p > .05$.

Discussion

Results suggested that across three MTurk samples, the demographic characteristics of the workers closely approximated those of the general US population with respect to gender and race. However, for religious affiliation, there were a very high number of Atheists and Agnostics (38.3%) found among MTurkers when compared to the general US population (4% in Lugo et al., 2008; 7.1% in Pew Research Center, 2014). The number of Atheists and Agnostics was consistently high in the three sets of data (29.4%, 37.7%, and 44.9%, respectively). These findings confirmed the results of previous research (Berinsky et al., 2012; Lewis et al., 2015). A high number of Agnostics and Atheists could be of concern to researchers studying religious and spiritual constructs who choose to use MTurk for their data collection. With such a high concentration of Agnostics and Atheists among MTurkers, those administering spiritual and religious measures might find their results skewed by the overrepresentation of this group in the dataset which may result in restriction of range in scores on these measures.

In terms of sample repeatability, significant differences were found between groups in terms of age, race, religious affiliation, measures of personality, and measures of spirituality and religiousness between MTurk samples obtained two days and six months apart. Overall sample means for the personality measure were within normative range, except for Extraversion ($M = 43.69$) which was lower (M T-score range: 45–55). Persons who make up a crowdsourcing sample might be more introverted and are therefore drawn to the solitary comforts of online interaction. Researchers using the Five Factor Model personality constructs will need to actively anticipate this seemingly endemic quality of MTurkers (see also Goodman et al., 2012).

Overall sample means for other scales on spirituality were found to be in the average range in comparison to the normative data (Piedmont, 2010), except that the Religious Involvement Scale scored very low ($M = 41.81$) in all three sets of data, no doubt a result of the higher incidence of Atheists and Agnostics sampled. These sampling differences may be of concern to future researches planning to utilise MTurk sample pools for conducting spiritual and religious inquiries in replicating research findings based on data from non-MTurk samples. However, the effect sizes noted were all in the small range (Cohen, 1992), suggesting that more research is necessary to determine whether these observed significant cohort differences represent inherent levels of sampling variability in the MTurk worker group or are merely an artefact of the rather large samples examined.

What is clear is that the consistent patterns of scores and demographics noted across studies were not due to the samples including the same workers. Each of the three samples can be considered relatively independent groups, making these observed patterns a function of the MTurk worker pool and not a result of the systematic inclusion of a particular subset of workers.

While change across cohorts was the norm, it does need to be pointed out that we did not identify any seasonal effects. Sample 3, which was obtained about six months after Samples 1 and 2, was not systematically different from these earlier cohorts in terms of

their lower levels of Extraversion, Religious Involvement, and the proportion of Atheists and Agnostics. Future researchers may want to examine more specifically whether the personality types involved in the worker pool vary as a function of time of year.

The results of this study suggested that sample variability was inherent in the MTurk population. We found that while there were significant changes across the three data sets, the effect sizes were small (eta^2 ranged from .01 to .04). Therefore, researchers need to determine whether this kind of variability will confound inferences drawn from their data. The most important finding was the consistencies in both the observed changes and stable scale profiles across the three studies. In a small sample, researchers should carefully examine who is involved in the study, because there may be gaps in the demographic composition of the participants. In large samples, these differences may lead to erroneous conclusions regarding mean levels on measured variables. Particularly individuals working in the fields of personality and spirituality need to be aware that using a MTurk sample might build up slight biases in their data (e.g., restriction of range for scores on certain scales). Researchers may need to be more proactive in recruiting people and in articulating the criteria they use for selecting people. For example, those working with spiritual and religious construct may wish to actively select only individuals who are involved in organised religions.

While mean levels varied across the measured variables, the underlying relations among these variables did not significantly differ across samples. That the fundamental structure of these variables was similar gives researchers confidence in both the comparability of data when wishing to aggregate findings from multiple samples, and in the robustness of findings obtained from any single study. Given that Box's *M* test is notoriously sensitive with large samples (Tabachnick & Fidell, 2007), its lack of significance in the current study, which employed three large samples, underscores the structural robustness of the measured variables. Further research will need to examine structural consistency when smaller samples are employed.

A limitation of the current study was that workers were restricted to the US population. Because MTurkers are shifting to an international audience (the second highest MTurk population is found in India), future research may want to broaden their scope to include non-US workers to get a more robust idea of who MTurkers are. Also, the range of constructs examined was relatively limited, centering mostly on personality and spirituality/religiousness. It would be important to extend this type of analysis to include other relevant individual difference constructs such as cognitive and behavioral indices. While this study did identify cohort differences, it did not attempt to determine the sources of these differences. Because MTurkers complete these surveys under non-controlled circumstances, they may be more prone to encounter distractions in their home environments that may impact their scores (e.g., cell phone activity or multitasking during the survey; Chandler, Mueller, & Paolacci, 2014; Clifford & Jerit, 2014; Goodman, et al., 2012). This issue deserves greater attention.

In conclusion, MTurk provides a number of advantages to researchers interested in collecting large, diverse samples cost effectively and quickly. However, attention does need to be given to the representativeness of those samples. Researchers may need to be more proactive and specific in their recruitment statements in order to ensure that responses to items reflect a full range of values. While some differences in sample characteristics between MTurk samples and other groups (e.g., college students) have already been

noted, the current study demonstrated that there are some stable, non-random qualities that characterise MTurk samples that researchers will need to be mindful of when conducting studies that employ personality and spiritual/religious constructs. Future research will need to explore further if there are any external factors which may lead to other non-random distortions in obtained samples (e.g., seasonal effects).

Acknowledgements

The authors would like to thank Rose Piedmont and Drs Martin F. Sherman and Joanna I. Hayward for their comments on an earlier version of this paper.

Disclosure statement

Ralph L. Piedmont is the author and publisher of the ASPIRES. He receives royalties from its sale.

ORCID

Ralph L. Piedmont http://orcid.org/0000-0002-0482-6761

References

Berinsky, A., Huber, G., Lenz, G., & Alvarez, R. M. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351–368. doi:10.1093/pan/mpr057

Bohannon, J. (2016). Mechanical Turk upends social science. *Science*, *352*, 1263–1264. doi:10.1126/science.352.6291.1263

Bradburn, N. M. (1969). *The structure of psychological well-being*. Chicago, IL: Aldine.

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*, 3–5. doi:10.1177/1745691610393980

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

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*, 120–131. doi:10.1017/xps.2014.5

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Crumbaugh, J. (1968). Purpose-in-Life Test. *Journal of Individual Psychology*, *24*, 74–81. doi:10.1126/science.352.6291.1263

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, *75*, 47–52. doi:10.1016/j.paid.2014.11.017

Frankl, V. E. (1985). *Man's search for meaning*. New York, NY: Pocket Books.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, *4*, 26–42. doi:10.1037/1040-3590.4.1.26

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224. doi:10.1002/bdm.1753

Guttmann, D. (1996). *Logotherapy for the helping professional*. New York, NY: Springer.

Horton, J., & Chilton, L. (2010). *The labor economics of paid crowdsourcing*. *Proceedings of the 11th ACM Conference on Electronic Commerce*. Retrieved from https://arxiv.org/pdf/1001.0627.pdf

Ipeirotis, P. (2012). *Demographics of Mechanical Turk*. Retrieved from http://www.ipeirotis.com/wp-content/uploads/2012/02/CeDER-10-01.pdf

Kam, C. D., Wilking, J. R., & Zechmeister, E. J. (2007). Beyond the 'narrow data base': Another convenience sample for experimental research. *Political Behavior*, *29*, 415–440. doi:10.1007/s11109-007-9037-6

Lewis, A. R., Djupe, P. A., Mockabee, S. T., & Su-Ya Wu, J. (2015). The (non) religion of Mechanical Turk workers. *Journal for the Scientific Study of Religion*, *54*, 419–428. doi:10.1111/jssr.12184

Lowenthal, M. F., Thurner, M., & Chiriboga, D. (1975). *Four stages of life: A psychological study of men and women facing transitions*. San Francisco, CA: Jossey-Bass.

Lugo, L., Stencel, S., Green, J., Smith, G., Cox, D., Pond, A., & Taylor, P. (2008). *US religious landscape survey: Religious affiliation-diverse and dynamic*. Retrieved from http://www.pewforum.org/files/2013/05/report-religious-landscape-study-full.pdf

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1–23. doi:10.3758/s13428-011-0124-6

Michel, J. S., O'Neill, S. K., Hartman, P., & Lorys, A. (2017). Amazon's Mechanical Turk as a viable source for organizational and occupational health research. *Occupational Health Science*, 1–16. doi:10.1007/s41542-017-0009-x

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419. Retrieved from http://journal.sjdm.org/10/10630a/jdm10630a.pdf

Pew Research Center. (2014). *Religious landscape study*. Washington, DC: Author.

Piedmont, R. L. (2010). *Assessment of spirituality and religious sentiments, technical manual* (2nd ed.). Timonium, MD: Author.

Roberts, T. M., Jangha, A., Piedmont, R. L., Sherman, M. F., & Williams, J. E. G. (2015). *Factor structure and personality disorder correlates of responses to the 50-item IPIP Big Five Marker Scale*. Unpublished manuscript.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education, Inc.

US Bureau of Labor Statistics. (2014). *Labor force statistics from the Current Population Survey*. Retrieved from http://www.census.gov/quickfacts/table/PST045215/00