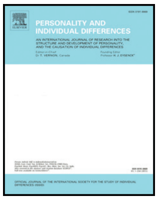
Personality and Individual Differences 133 (2018) 41–46

Contents lists available at ScienceDirect

Personality and Individual Differences

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



Exploring beyond simple demographic variables: Differences between traditional laboratory samples and crowdsourced online samples on the Big Five personality traits☆,☆☆

Douglas E. Colman [a,*], Jared Vineyard [b], Tera D. Letzring [a]

[a] *Department of Psychology, Idaho State University, United States*
[b] *Idaho Center for Health Research, Idaho State University, United States*

ARTICLE INFO

Article history:
Received 30 December 2016
Received in revised form 23 May 2017
Accepted 15 June 2017
Available online 20 June 2017

Keywords:
Crowdsourcing
Mechanical Turk
Personality
Individual differences
Big Five traits

ABSTRACT

Amazon's Mechanical Turk (MTurk), a popular crowdsourcing website, is increasingly being utilized by researchers to obtain psychological data. This transition has prompted evaluation of sourcing costs, psychometric properties, and motivations of participants. However, research is limited comparing traditional and crowdsourced participants on personality measures. Therefore, in the current study laboratory participants (drawn from three universities) and MTurk workers completed the Big Five inventory and provided demographic information using web-based surveys. Controlling for age and gender, laboratory participants were significantly lower in Openness ($\hat{d} = 0.26$), and higher in Extraversion ($\hat{d} = 0.37$), Agreeableness ($\hat{d} = 0.15$), and Neuroticism ($\hat{d} = 0.05$) than MTurk participants. However, pairwise comparisons among individual sites revealed there were means above and below that for MTurk participants for Openness and Conscientiousness. Given these differences, researchers are encouraged to consider how such personality characteristics may influence the outcomes of their research when designing and conducting psychological studies that use crowdsourcing techniques to recruit participants.

It has been said that what is done alone is sometimes better accomplished in a crowd. Modern behavioral researchers are now turning more frequently to cost-effective online solutions to sample and collect data from human participants (Howe, 2006). What is collectively known as crowdsourcing has become a popular avenue for data collection, with Amazon's Mechanical Turk (MTurk) becoming one of the most commonly used services. Crowdsourcing offers a significant advancement in the study of large populations (Paolacci & Chandler, 2014), but questions remain concerning the importance of individual differences between the more traditional laboratory samples in which data are collected in-person in a laboratory and MTurk samples in which data are collected online without participants coming to a laboratory. In addition, recent research has demonstrated that relatively wide variation exists in personality characteristics when participants are assessed with traditional in-person data collection across 30 colleges and universities (Corker, Donnellan, Kim, Schwartz, & Zamboanga, 2015). The current study explored differences across settings by comparing a large sample of MTurk participants to several college student samples that used in-person data collection.

Crowdsourcing refers to using the internet to distribute tasks or work among a large group of individuals for compensation (Behrend, Sharek, Meade, & Wiebe, 2011; Chandler & Shapiro, 2016; Howe, 2006). MTurk provides a platform for researchers to post tasks for “workers” to complete for relatively little compensation. Compensation ranges from a few cents for short studies up to several dollars, although the majority of tasks posted by researchers fall below $1.00. Recently, psychologists have increasingly used the internet, and crowdsourcing in particular, to recruit samples for studies and experiments traditionally gathered from community or university samples (Chandler & Shapiro, 2016; Skitka & Sargis, 2006).

The use of crowdsourcing websites provides a significant advantage to researchers in that large samples can be collected in a relatively short amount of time (Buhrmester, Kwang, & Gosling, 2011), compared to months (and sometimes years!) for data collection occurring in-person. In short, crowdsourcing is more efficient because it does not require physical lab space, eliminates the need for data entry, and allows for data collection at any time of the day or week. However, there has been concern over the equivalence of data quality among the various

☆ Portions of this research were presented at the 4th Biennial Association for Research in Personality Conference in St. Louis, MO in 2015.

☆☆ We thank Jeremy Biesanz and Simine Vazire for sharing data which comprise two of the samples reported in this article. We also thank Dolan Ward for his assistance with data collection and early contributions to this manuscript.

* Corresponding author at: Idaho State University, Department of Psychology, 921 S. 8th Ave, Stop 8112, Pocatello, 83209, United States.
E-mail address: colmdoug@isu.edu (D.E. Colman).

http://dx.doi.org/10.1016/j.paid.2017.06.023
0191-8869/

participant recruitment and data collection methods (Buhrmester et al., 2011; Gosling, Vazire, Srivastava, & John, 2004; Ward, 1993).

It has been argued that platform (paper-and-pencil, lab computer, and crowdsourced) differences could significantly alter the results of a study. Yet, evidence from several studies has demonstrated that differences between traditional in-person and crowdsourced participants are minimal. At the assessment level, there is already strong evidence of measurement equivalence/invariance for personality measures taken by MTurk workers and traditional in-person participants. Specifically, the 100 item IPIP version (Goldberg et al., 2006) of the NEO-PI-R was found to have equivalent psychometric properties across samples (Behrend et al., 2011). Likewise, the Big Five inventory (John, Naumann, & Soto, 2008) showed measurement invariance when MTurk participants were restricted to being from countries in which English is the primary language (Feitosa, Joseph, & Newman, 2015). In addition to efficiency and data equivalency, MTurk participants tend to be more diverse on important demographic variables such as age, education, and ethnicity than traditional in-person college/university participants (Behrend et al., 2011; Paolacci & Chandler, 2014). This greater demographic variability directly addresses one of the common limitations of studies conducted in-person at a single location.

Although there are benefits to crowdsourcing psychological data, such samples may not always produce conclusions that can be generalized to a non-MTurk population. One characteristic of crowdsourced participants that has the potential to inhibit generalizability is personality. Previous research has found that crowdsourced samples, in comparison to in-person samples, were lower on Extraversion, Neuroticism, Openness to experience, and Conscientiousness[1] when the Big Five traits were assessed with a 10-item inventory, as well as lower on trait level self-esteem when assessed with a single item (Goodman, Cryder, & Cheema, 2013; Kosara & Ziemkiewicz, 2010). A second study found that a crowdsourced sample again revealed lower levels of Extraversion and Neuroticism, but found higher levels of Openness and lower levels of Agreeableness (Kosara & Ziemkiewicz, 2010). Additionally, personality traits have also been found to be quite variable across traditional samples collected from different regions of the US (Corker et al., 2015), and therefore it is not yet clear whether the magnitude of the differences found between MTurk and traditional samples is within the range that would be expected based on regional differences.

Furthermore, it is well-known that differences in personality are related to behavioral differences that can result in variation in important life outcomes (Ozer & Benet-Martinez, 2006; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007), and therefore meaningful differences in personality between crowdsourced and traditional samples could be muddying the conclusions that can be drawn about basic psychological processes. Therefore, if consistent differences in personality exist between crowdsourced and traditional participants, it would behoove researchers to uncover and consider such differences when using crowdsourced samples.

Measuring and reporting variance in personality patterns across settings is a key aspect of understanding the impact of these differences for research efforts. Samples using crowdsourcing methods, university students, or community members differ on important variables. Therefore, replication of previous work and comparing multiple samples should be employed to strengthen the understanding of these differences. Using multiple samples allows for improved comparison within traditional settings and improves reliability of estimates of personality and related individual differences. Previous work has shown variability between crowdsourced and traditional samples (Goodman et al., 2013; Kosara & Ziemkiewicz, 2010), however a multiple-sample strategy is needed from both crowdsourced and traditional milieu to better examine the pattern of differences.

Table 1
Demographic characteristics by sample.

	Sample					
	A	B	C	D1	D2	D3
Gender						
Male	389	145	137	31	39	36
Female	880	355	279	135	88	79
Age						
M	37.15	21.77	19.43	21.00	19.50	20.41
SD	12.55	5.32	2.32	4.09	2.02	3.43
Range	18–78	18–53	18–39	17–50	17–35	17–44

1. Summary

Recent research has noted significant variability in participant personality across the United States and concluded that studies comparing a limited number of participant sources run the risk of over-generalizing research findings (Corker et al., 2015). On the other hand, data from crowdsourced participants are thought to provide high quality data (Buhrmester et al., 2011) and incorporate a wider range of individuals with regard to age and ethnic background (Behrend et al., 2011). Less clear, however, is the personality profile of crowdsourced participants compared to traditional in-person participants. Thus, the current analyses expand upon the framework of Corker et al. (2015) by examining the individual differences among participants from several in-person samples and a larger sample of crowdsourced participants (Table 1).

2. Method

2.1. Participants

2.1.1. Sample A

These data ($N = 1279$) were collected between Fall 2014 and the end of 2015 via an online survey platform. Data were collected as part of several projects for which all participants were recruited from MTurk, but restricted to individuals with an approval rating of 95% or greater who reside in the United States.[2] Participants were financially compensated between $0.50 and $2.00, depending upon the project.

2.1.2. Sample B

These data ($N = 500$) were collected between Fall 2014 and Spring 2016 in person at Idaho State University located in the West region of the United States. Data were collected as part of two separate projects for which all participants were recruited from a department subject pool. Announcements about each project were also made in some courses where research was a required element. As such, all participants were undergraduate students and were remunerated with research credits.

2.1.3. Sample C

These data ($N = 418$) were collected between Fall 2012 and the end of 2013 in person at Washington University in St. Louis, a private university in the Midwest region of the United States (Vazire et al., 2016). These participants were taking part in a longitudinal study for which they receive financial compensation. Recruitment occurred through various methods, including a department subject pool, flyers, and announcements in classes. Most of the participants were undergraduate students but a few (about 7%) were graduate students.

2.1.4. Sample D

These data ($N = 408$) were collected in or after 2007 and prior to 2014 in person at University of British Columbia located in British Columbia, Canada. Data were collected as part of several projects in

[1] Differences for Extraversion, Neuroticism, and self-esteem were found in two studies, while differences for Conscientiousness and Openness were only found in one study.

[2] All participants met inclusion criteria, which consisted of successfully answering ≥80% of embedded attention checks and completing ≥80% of the procedure in the given study.

which participants received financial compensation or were remunerated with course credit. This dataset was compiled for a previously published study (Rogers & Biesanz, 2014), but the results presented herein are not previously published. Because of the ethnic diversity of undergraduate students in this sample, we utilized the same three ethnic sub-groupings outlined in Rogers and Biesanz (2014). These subgroups are Euro-Canadian (Sample D1; $n = 166$), Acculturated East Asian (Sample D2; $n = 127$), and Semi-Acculturated East Asian (Sample D3; $n = 115$).

2.2. Measures

2.2.1. Big Five inventory

The self-report version of the 44-item Big Five inventory (BFI; John et al., 2008) was used to assess participants' personality trait dimensions of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Respondents rated the applicability of short phrases (e.g., *does a thorough job* for Conscientiousness and *is original, comes up with new ideas* for Openness) on a Likert Scale. This measure has been demonstrated to have adequate reliability with Cronbach's alpha coefficients from 0.79 to 0.88 for the five subscales and 0.83 for the overall measure (Benet-Martínez & John, 1998). The minimum Cronbach's alpha for self-reports across samples were 0.74, 0.65, 0.87, 0.77, and 0.82 for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, respectively. Noteworthy is that the width of the Likert Scales varied across the samples; some samples used a 5-point width, while one implemented a 7-point width, and another a 15-point width. Thus, to make the scaling comparable across samples, subscale scores were converted using the *Percentage of Maximum Possible* (POMP) method (Cohen, Cohen, Aiken, & West, 1999). This conversion was completed using the following equation: $\left[\frac{\textit{Observed scale score}-\textit{ Min. possible scale score})}{\textit{Max. possible scale score}-\textit{ Min. possible scale score})}\right] \times 100$. As such, values reported herein have a real range from 0 to 100.

3. Results

The goal of this study was to discover how the Big Five personality characteristics vary across samples, with special consideration toward crowdsourcing. First, however, we compared estimates of the Big Five traits in our study to those estimated in recent research. Specifically, collapsed across in-person samples, the observed trait levels are similar to those observed by Corker et al. (2015) across 30 college/university samples for Conscientiousness and Extraversion, lower for Openness and Agreeableness, and higher for Neuroticism.[3] The point estimates and their 95% confidence intervals are listed in Table 2.

3.1. Are crowdsourced participants similar across geographic location?

In a similar vein to the study by Corker et al. (2015), we explored geographic differences in the personality trait levels of crowdsourced participants. For a large proportion (81.1%; 1037 of 1279) of MTurk participants (those from Sample A), state of residence was reported. This subsample contained individuals from every U.S. state as well as Washington D.C. However, this subsample was too small (e.g., two states were only reported once) to compare personality trait levels across all states. As such, a larger, generally accepted grouping criterion was used. Specifically, location was grouped into the four regions (Northeast, Midwest, South, and West) and nine divisions (New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific) outlined by the United States Census Bureau (n.d.).

Table 2
Mean values and 95% confidence intervals for in-person participants in Corker et al., 2015 and current sample.

Personality trait	Corker et al., 2015			Current sample		
	M	95% CI		*M*	95% CI	
		Lower	Upper		Lower	Upper
Openness	68.75	67.75	69.50	64.45	63.68	65.21
Conscientiousness	64.25	63.50	65.75	64.31	63.46	65.17
Extraversion	57.75	56.25	59.00	58.20	57.15	59.25
Agreeableness	73.25	72.50	75.00	70.08	69.28	70.89
Neuroticism	45.00	44.00	46.00	48.07	47.04	49.10

Note. The values listed from the Corker et al. (2015) paper are the meta-analytic estimates across sites (e.g., grand means) presented in Fig. 1 of the published paper. To make the scales directly comparable, we transformed the values using the POMP method (Cohen et al., 1999).

We first examined whether participants were similar demographically across the geographic locations. Across the four regions and nine divisions, there were no differences in the frequencies of reported gender ($\chi^2(3) = 0.96, p = 0.81$; $\chi^2(8) = 3.52, p = 0.90$, respectively) or age ($F(3, 1025) = 1.12, p = 0.34, \eta_p^2 = 0.003$; $F(8, 1020) = 1.54, p = 0.14, \eta_p^2 = 0.012$, respectively).

Next, we examined regional differences on the Big Five personality traits across geographic locations. The Big Five traits are theoretically orthogonal (e.g., Costa & McCrae, 1995; Goldberg, 1993), and therefore differences across locations were explored using univariate analyses. In the current subsample, there were no significant differences across the four regions for Conscientiousness, Extraversion, and Neuroticism (all $F < 0.60, p > 0.61, \eta_p^2 < 0.002$). There were, however, significant differences for Openness ($F(3, 1033) = 2.64, p = 0.05, \eta_p^2 = 0.008$) and Agreeableness ($F(3, 1032) = 2.73, p = 0.04, \eta_p^2 = 0.009$). Tukey post hoc analyses for Openness revealed that none of the differences reached statistical significance, with the largest difference observed between the West and Midwest regions ($M_{\text{diff}} = 0.15$, 95% CI $[-0.01, 0.32]$,[4] $p = 0.09, \hat{d} = 0.22$).[5] Alternatively, the Tukey post hoc analyses for Agreeableness revealed that only the South and Northeast regions were significantly different ($M_{\text{diff}} = 0.17$ $[0.003, 0.34]$, $p = 0.04, \hat{d} = 0.24$).

Lastly, differences across the nine divisions were also explored using univariate analyses. These analyses indicated there were no significant differences across the nine divisions for Conscientiousness, Extraversion, Agreeableness, and Neuroticism (all $F < 1.49, p > 0.15, \eta_p^2 < 0.011$). However, there were significant differences in Openness ($F(8, 1028) = 2.50, p = 0.01, \eta_p^2 = 0.02$). Examination of the Tukey post hoc analyses revealed that only two regions had a statistically significant difference – the Pacific and Middle Atlantic divisions ($M_{\text{diff}} = 0.25$ $[0.004, 0.50]$, $p = 0.04, \hat{d} = 0.37$).[6]

3.2. Are crowdsourced participants different than traditionally sourced participants?

We first explored age and gender differences. Using an independent-samples t-test, it was found that, in line with previous research (Behrend et al., 2011; Paolacci & Chandler, 2014), participants who were recruited to participate in person were significantly younger ($M = 20.61, SD = 4.12$) than the crowdsourced sample ($M = 37.15, SD = 12.55$; $t(1520) = 44.53, p < 0.001$, 95% CI of the difference

[3] Corker et al. (2015) used the Mini IPIP (Donnellan, Oswald, Baird, & Lucas, 2006) to assess the Big Five personality domains.

[4] Values in brackets denote 95% confidence intervals for the remainder of the results.

[5] Cohen's d effect sizes were estimated using the following formula for all pairwise post-hoc comparisons: $\hat{d} = \frac{\overline{X'_i} - \overline{X'_k}}{\sqrt{MS'_{error}}}$.

[6] Our MTurk sample was composed of data collected as part of two different projects which took place approximately one year apart (Fall 2014 vs. Fall 2015). As such, we conducted a series of t-tests on the Big Five personality traits between these two time points. Of these analyses, we only found a significant difference in Conscientiousness ($t(356.02) = 2.81, p = 0.005, M_{\text{diff}} = 3.59$ $[1.08, 6.09]$, $d = 0.19$).

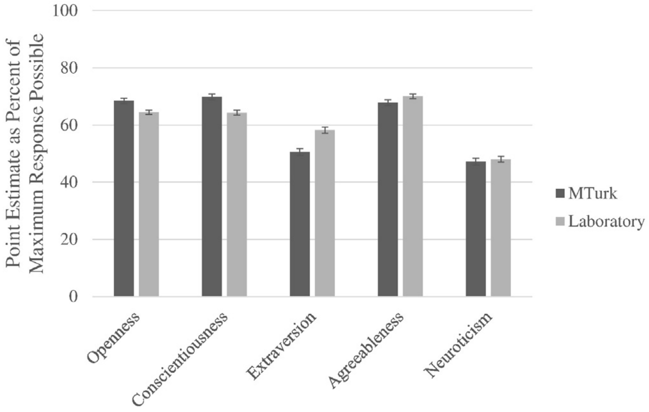


Fig. 1. Unadjusted mean values on the Big Five personality traits for in-person Laboratory participants (combined across samples) and Amazon's Mechanical Turk (MTurk) participants. Displayed means and 95% confidence interval are shown in the Percentage of Maximum Possible (Cohen et al., 1999) metric, which has a real range from 0 to 100.

[15.81, 17.27], $d = 1.78$).[7] Additionally, chi-square tests were used to examine the distribution of gender within and between samples. There were disproportionally more females than males overall ($\chi^2(1) = 416.32$, $p < 0.001$). Additionally, based on a 2 (gender) × 2 (sample) chi-square test of association, this difference in the number of female and male participants was consistent for both samples; $\chi^2(1) = 0.50$, $p = 0.48$. Given these findings, some researchers may advocate for using these variables as covariates when examining differences across samples. However, heeding the recommendations by Simmons, Nelson, and Simonsohn (2011), we have opted to report analyses with and without the covariates of age and gender.

Independent-samples *t*-tests were used to compare MTurk and traditional in-person participants on each of the Big Five traits. Compared to MTurk participants, in-person participants had lower levels of Openness ($t(2482.3) = 6.52$, $p < 0.001$, $M_{\text{diff}} = 4.00$, [2.80, 5.21], $d = 0.26$) and Conscientiousness ($t(2511.4) = 8.14$, $p < 0.001$, $M_{\text{diff}} = 5.51$, [4.18, 6.84], $d = 0.32$), higher levels of Extraversion ($t(2525.4) = 9.26$, $p < 0.001$, $M_{\text{diff}} = 7.62$, [6.01, 9.23], $d = 0.36$) and Agreeableness ($t(2471.5) = 3.54$, $p < 0.001$, $M_{\text{diff}} = 2.31$, [1.03, 3.59], $d = 0.14$), and similar levels of Neuroticism ($t(2540.4) = 1.11$, $p = 0.27$, $M_{\text{diff}} = -0.88$, [−2.44, 0.68], $d = 0.04$). Inclusion of age and gender as covariates in a series of ANCOVAs produced similar, but not fully parallel, results. These models also indicated that in-person participants possessed lower levels of Openness x($F(1, 2551) = 19.59$, $p < 0.001$, $\hat{d} = 0.26$) and higher levels of Extraversion ($F(1, 2551) = 66.29$, $p < 0.001$, $\hat{d} = 0.37$) and Agreeableness ($F(1, 2550) = 41.23$, $p < 0.001$, $d = 0.15$). However, Neuroticism ($F(1, 2548) = 21.36$, $p < 0.001$, $\hat{d} = 0.05$) was higher for in-person participants while Conscientiousness had similar levels to MTurk participants ($F(1, 2551) = 0.07$, $p = 0.80$, $\hat{d} = 0.32$). See Fig. 1 for the unadjusted means and 95% confidence interval for each personality trait by recruitment method.

3.3. Assessment of the Big Five personality traits across all samples

With only a single pairwise difference between regions and a single pairwise difference between divisions in trait levels, we conducted the following analyses under the assumption that MTurk participants constitute a single recruitment and data collection sample. Therefore, trait levels were compared across six groups: MTurk and each of the five traditional in-person samples. We again conducted the analyses both without age and gender as covariates (using ANOVAs) and with age and gender as covariates (using ANCOVAs). Without controlling for age and gender, the analyses revealed that sample means were significantly different for the traits of Openness ($F(5, 2593) = 17.32$, $p < 0.001$, $\eta_p^2 = 0.03$), Conscientiousness ($F(5, 2593) = 43.08$, $p < 0.001$, $\eta_p^2 = 0.08$), Extraversion ($F(5, 2593) = 18.10$, $p < 0.001$, $\eta_p^2 = 0.03$), and Agreeableness ($F(5, 2592) = 9.67$, $p < 0.001$, $\eta_p^2 = 0.02$), although the effect sizes were small. However, there were no significant differences across samples for the trait of Neuroticism ($F(5, 2590) = 1.92$, $p = 0.09$, $\eta_p^2 = 0.004$).

When controlling for age and gender, the results were again similar to the analyses that did not control for age and gender, but not a full parallel. Specifically, there were significant differences across samples for each Big Five trait: Openness, $F(5, 2547) = 12.21$, $p < 0.001$, $\eta_p^2 = 0.02$; Conscientiousness, $F(5, 2547) = 26.92$, $p < 0.001$, $\eta_p^2 = 0.05$; Extraversion, $F(5, 2547) = 14.08$, $p < 0.001$, $\eta_p^2 = 0.03$; Agreeableness, $F(5, 2546) = 14.17$, $p < 0.001$, $\eta_p^2 = 0.03$; and Neuroticism, $F(5, 2544) = 5.94$, $p < 0.001$, $\eta_p^2 = 0.01$. The unadjusted means and 95% confidence interval for each sample are displayed by personality trait in Fig. 2. Further, the adjusted effect sizes for the pairwise comparisons are listed in Table 3 for each type of model (no covariates vs. covariates) for each trait.

As seen in Table 3, there were many significant differences among the samples with regards to personality traits. For Openness and Conscientiousness, there is a large amount of variability among all the samples, although MTurk participants tended to be somewhat higher on these two traits. On the contrary, there was little difference across in-person samples for Extraversion and Agreeableness. However, MTurk participants were consistently lower than all other samples on these two traits. Lastly, and interestingly, there was little difference across all samples on Neuroticism.

4. Discussion

Sampling is a critical element to the external validity of studies (Landers & Behrend, 2015). Specifically, sampling strategy can limit generalizability through two means – range restriction and omitted variable bias. Traditional in-person participant data sourcing often consists of college students with a limited range of age, life experience, and ethnic diversity. Thus, one advantage to crowdsourcing psychological data is the diversity of the participants (Behrend et al., 2011; Mason & Suri, 2012; Paolacci & Chandler, 2014). For example, in the current MTurk sample an increased range and variability for age was found compared to the traditionally sourced participants (range = 18 to 78 for MTurk vs. 17 to 53 for in-person samples; $SD = 12.52$ for MTurk vs. 5.32 for the most variable in-person sample).

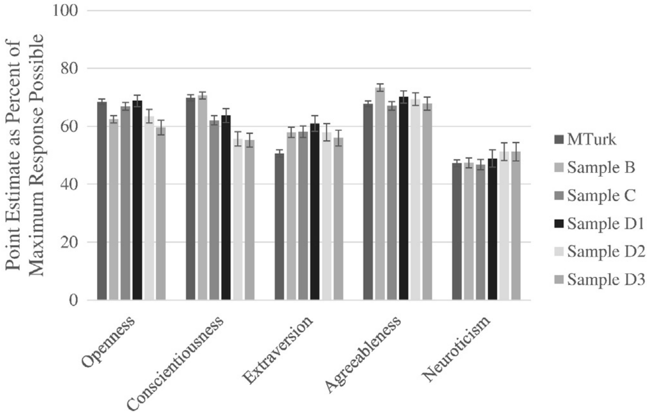


Fig. 2. Unadjusted mean values on the Big Five personality traits for each data collection sample. Displayed means and 95% confidence interval are shown in the Percentage of Maximum Possible (Cohen et al., 1999) metric, which has a real range from 0 to 100.

[7] Following the recommendations of Delacre, Lakens, and Leys (2017) we defaulted to Welch two sample *t*-tests for analyses. Therefore, the reported degrees of freedom are adjusted.

Table 3
Estimated Cohen's *d* effect sizes for the pairwise comparisons across samples by trait.

Sample vs. sample	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
A vs. __										
B	0.39^{***}	0.38^{***}	−0.05	−0.05	-0.35^{***}	-0.35^{***}	-0.34^{***}	-0.34^{***}	−0.01	−0.01
C	0.10	0.11	0.46^{***}	0.45^{***}	-0.36^{***}	-0.36^{***}	0.04	0.04	0.02	0.01
D1	−0.02	−0.02	0.37^{***}	0.38^{***}	-0.50^{***}	-0.50^{***}	−0.14	−0.14	−0.08	−0.09
D2	0.32^{**}	0.31^{*}	0.84^{***}	0.86^{***}	-0.35^{**}	-0.35^{**}	−0.09	−0.10	−0.20	−0.20
D3	0.57^{***}	0.57^{***}	0.87^{***}	0.89^{***}	−0.26	−0.26	0.00	0.00	−0.20	−0.21
B vs. __										
C	-0.29^{***}	-0.27^{***}	0.51^{***}	0.50^{***}	−0.01	−0.01	0.38^{***}	0.38^{***}	0.03	0.03
D1	-0.41^{***}	-0.41^{***}	0.41^{***}	0.43^{***}	−0.15	−0.15	0.19	0.20	−0.08	−0.07
D2	−0.07	−0.07	0.89^{***}	0.91^{***}	0.00	0.00	0.24	0.25	−0.19	−0.19
D3	0.18	0.18	0.92^{***}	0.94^{***}	0.09	0.09	0.33^{*}	0.34^{*}	−0.19	−0.19
C vs. __										
D1	−0.12	−0.13	−0.10	−0.08	−0.14	−0.14	−0.18	−0.18	−0.10	−0.10
D2	0.21	0.20	0.38^{**}	0.41^{***}	0.01	0.01	−0.14	−0.14	−0.22	−0.22
D3	0.47^{***}	0.45^{***}	0.41^{**}	0.44^{***}	0.10	0.10	−0.04	−0.04	−0.22	−0.22
D1 vs. __										
D2	0.34	0.34	0.48^{***}	0.49^{***}	0.15	0.15	0.05	0.05	−0.12	−0.12
D3	0.59^{***}	0.59^{***}	0.50^{***}	0.51^{***}	0.24	0.24	0.14	0.14	−0.12	−0.12
D2 vs. __										
D3	0.25	0.25	0.03	0.03	0.09	0.09	0.09	0.10	0.00	0.00

Note. The left column for each trait contains the values derived from the model without covariates, while the right column for each trait contains the values derived from the model with the covariates of age and gender included. Cohen's *d* effect sizes were estimated using the following formula: $\hat{d} = \frac{\overline{X'_i} - \overline{X'_k}}{\sqrt{MS'_{error}}}$.

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

In regards to the focus of the present article, some research suggests that personality varies significantly between crowdsourced samples and traditional community and college student samples (Goodman et al., 2013; Kosara & Ziemkiewicz, 2010). In the current study in which traditional samples from multiple locations were used, a similar pattern of results was found. Additionally, these findings generally held even after controlling for the covariates of age and gender, a step not taken in previous work. Specifically, the MTurk workers were more Open to Experience, but less Extraverted, Agreeable, and Neurotic compared to traditionally recruited participants.

Issues might arise, however, when trying to make inferences about the MTurk population by comparing a large sample of workers to traditionally sourced participants (e.g., students, community members). For instance, research has recently shown that traditional samples vary significantly in the Big Five personality characteristics across locations (Corker et al., 2015). Thus, we first examined differences in the personality characteristics of MTurk workers across regions and divisions of the U.S. Interestingly, unlike traditional samples, there was only one significant pairwise difference based on geographic region across the five personality traits. Similarly, there was only a single significant pairwise difference across the nine U.S. divisions. Given these findings, we felt it justified to treat the MTurk workers stemming from the various geographic locations within the U.S. as a single sample as we explored differences between MTurk workers and the traditional, in-person research participants.

In a similar vein, generalizability could be inhibited if only a single sample of in-person participants was compared to the MTurk sample. Thus, in addition to the large MTurk sample, three different traditional samples in which data collection location varied widely were used to explore differences in personality. From this, it was found that MTurk workers tended to have higher levels of Openness and Conscientiousness, but lower levels of Extraversion and Agreeableness. This is a similar pattern of results to the previous research (Goodman et al., 2013; Kosara & Ziemkiewicz, 2010) and the comparison between all in-person participants (combined across samples) and MTurk workers. However, we see from Table 2 that these are general patterns and *not absolute trends*. Thus, researchers should remain cognizant of differences among recruitment and data collection sites when sourcing psychological data – whether crowdsourcing techniques are being implemented or not.

While Landers and Behrend (2015) discuss the personality characteristics of convenience samples in the context of range restriction, our findings herein should fall under the purview of both range restriction and variable omission problems. Range would be more restricted in student samples, as they will be younger and thus have less life experience. For example, researchers interested in job, career, or management related constructs would likely find different results across sampling locations and methods for this reason. Indeed, significant differences between students and working adult samples have been found (Ward, 1993), which significantly limits the generalizability of traditional samples in such research domains.

On the other hand, omitting personality variables when considering relations is the cause for concern. Research on the relation between income and life satisfaction exemplifies this issue. Specifically, it has been reported that income is an important factor in one's life satisfaction (Frijters, Haisken-DeNew, & Shields, 2004). However, it also has been shown that higher Conscientiousness, Extraversion, and Agreeableness, and lower Neuroticism, predicted greater life satisfaction even when accounting for income (Soto & Luhmann, 2013). Furthermore, Neuroticism moderated this relationship in that life satisfaction was more strongly related to income for highly neurotic vs. emotionally stable individuals. These findings exemplify the fact that failing to account for differences in these types of characteristics among samples can, and often do, lead to over-generalized, and sometimes inappropriate, conclusions.

In light of the current findings and those in prior research, we strongly recommend future research studies to take individual differences, especially the Big Five personality traits, into account during design, collection, and analysis of crowdsourced data. Ultimately, if researchers do consider, and go as far as statistically controlling for, characteristics of participants that likely affect the constructs of interest, MTurk is a great way to obtain a more diverse sample that contains participants that hold a greater amount of life experiences. Such considerations represent an extremely important process in the generalization of results gleaned from convenience samples (Landers & Behrend, 2015), such as crowdsourced workers. Not only will these considerations allow researchers to appropriately generalize results, but it may lead to theory revision and even new discoveries, through the exploration of individual differences as potential mediators and/or moderators of the relationship being explored.

References

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavioral Research Methods*, *43*, 800–813. http://dx.doi.org/10.3758/s13428-011-0081-0.

Benet-Martínez, V., & John, O. P. (1998). *Los Cinco Grandes* across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, *75*, 729–750.

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. http://dx.doi.org/10.1177/1745691610393980.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*, 53–81. http://dx.doi.org/10.1146/annurev-clinpsy-021815-093623.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*, 315–346.

Corker, K. S., Donnellan, M. B., Kim, S. Y., Schwartz, S. J., & Zamboanga, B. L. (2015). College student samples are not always equivalent: The magnitude of personality differences across colleges and universities. *Journal of Personality*, *85*, 123–135. http://dx.doi.org/10.1111/jopy.12224.

Costa, P. T., Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the revised NEO personality inventory. *Journal of Personality Assessment*, *64*, 21–50.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, *30*. http://dx.doi.org/10.5334/irsp.82.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*, 192–203. http://dx.doi.org/10.1037/1040-3590.18.2.192.

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. http://dx.doi.org/10.1016/j.paid.2014.11.017.

Frijters, P., Haisken-DeNew, J. P., & Shields, M. A. (2004). Money does matter! Evidence from increasing real income and life satisfaction in East Germany following reunification. *The American Economic Review*, *94*, 730–740.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26–34.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84–96. http://dx.doi.org/10.1016/j.jrp.2005.08.007.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224. http://dx.doi.org/10.1002/bdm.1753.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93–104. http://dx.doi.org/10.1037/0003-066X.59.2.93.

Howe, J. (2006). *The rise of crowdsourcing*. *14*. (pp. 1–5). Wired Magizine, 1–5.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift in the integrative big five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York, NY: Guilford Press.

Kosara, R., & Ziemkiewicz, C. (2010). Do mechanical turks dream of square pie charts? *Paper presented at the 3rd BELIV'10 workshop: Beyond time and errors: Novel evaluation methods for information visualization, Atlanta, GA*.

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*, 142–164. http://dx.doi.org/10.1017/iop.2015.13.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's mechanical turk. *Behavior Research Methods*, *44*, 1–23. http://dx.doi.org/10.3758/s13428-011-0124-6.

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401–421. http://dx.doi.org/10.1146/annurev.psych.57.102904.190127.

Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188. http://dx.doi.org/10.1177/0963721414531598.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 313–345.

Rogers, K. H., & Biesanz, J. C. (2014). The accuracy and bias of interpersonal perceptions in intergroup interactions. *Social Psychological and Personality Science*, *5*, 918–926. http://dx.doi.org/10.1177/1948550614537307.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. http://dx.doi.org/10.1177/0956797611417632.

Skitka, L. J., & Sargis, E. G. (2006). The internet as psychological laboratory. *Annual Review of Psychology*, *57*, 529–555. http://dx.doi.org/10.1146/annurev.psych.57.102904.190048.

Soto, C. J., & Luhmann, M. (2013). Who can buy happiness? Personality traits moderate the effects of stable income differences and income fluctuations on life satisfaction. *Social Psychological and Personality Science*, *4*, 46–53. http://dx.doi.org/10.1177/1948550612444139.

United States Census Bureau. (n.d.). Geographic terms and concepts - Census divisions and census regions. Retrieved from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

Vazire, S., Wilson, R. E., Bollich, K. L., Solomon, B. C., Carlson, E. N., Harris, K., … Jackson, J. J. (2016). Personality and Interpersonal Roles Study (PAIRS). Unpublished data. Codebook available at https://osf.io/akbfj/.

Ward, E. A. (1993). Generalizability of psychological research from undergraduates to employed adults. *The Journal of Social Psychology*, *133*, 513–519.