

- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419.
- Rand, D. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299, 172–179.
- Ross, J., Irani, I., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Amazon Mechanical Turk. In K. Edwards & T. Rodden (Eds.), *Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 2863–2872). New York, NY: ACM.
- Shapiro, D., Chandler, J., & Mueller, P. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, 1, 213–220.

## Don't Throw the Baby Out With the Bathwater: Comparing Data Quality of Crowdsourcing, Online Panels, and Student Samples

Nicolas Roulin  
*University of Manitoba*

In their focal article, Landers and Behrend (2015) propose to reevaluate the legitimacy of using the so-called convenience samples (e.g., crowdsourcing, online panels, and student samples) as compared with traditional organizational samples in industrial–organizational (I-O) psychology research. They suggest that such sampling strategies should not be judged as inappropriate per se but that decisions to accept or reject such samples must be empirically or theoretically justified. I concur with Landers and Behrend's call for a more nuanced view on convenience samples. More precisely, I suggest that we should not “throw the baby out with the bathwater” but rather carefully and empirically examine the advantages and risks associated with using each sampling strategy before classifying it as suitable or not.

In this commentary, I examine and compare original data obtained from the three types of convenience samples highlighted in the focal article: (a) crowdsourcing, (b) online panels with participants recruited by commercial panel providers, and (c) business students. I discuss differences with regards to sample composition (i.e., age, gender, ethnicity, education, and

Nicolas Roulin, Asper School of Business, Department of Business Administration, University of Manitoba, Winnipeg, Manitoba, Canada.

I want to thank Steve Binggeli, Joshua S. Bourdage, Vanessa Jacksch, Ute-Christine Klehe, Franciska Krings, and Rima Tarraf for their help in the data collection, as well as Franciska Krings for her feedback on a draft of the article.

Correspondence concerning this article should be sent to Nicolas Roulin, Asper School of Business, Department of Business Administration, University of Manitoba, 406 Drake Center, Winnipeg, Manitoba, Canada R3T 5V4. E-mail: [nicolas.roulin@umanitoba.ca](mailto:nicolas.roulin@umanitoba.ca)

employment status). Moreover, using the example of a measure of competitive worldviews (CWs; Duckitt, Wagner, Du Plessis, & Birum, 2002), I examine data quality (i.e., mean differences and range restriction, scale reliability, data normality). I compare data from the original convenience samples to published data in the social and personality psychology literatures obtained with two more traditional (but still convenience) sample types: psychology students and survey data from participants recruited from the general adult population.

### Convenience Samples Composition

In the focal article, Landers and Behrend (2015) argue that online sampling strategies such as crowdsourcing (e.g., Mechanical Turk) may reduce research overreliance on WEIRD (Western, educated, industrialized, rich, and democratic) samples. Such strategies may be considered as a viable alternative to the traditional convenience samples composed of young and inexperienced college students, for instance when organizational samples are difficult to access. Indeed, Mechanical Turk (MTurk) participants have been described as more diverse than traditional college samples and more representative of the general (U.S.) population (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). However, there is limited literature discussing the composition of other forms of online panels (e.g., provided by commercial participant recruitment services like Qualtrics or SurveyMonkey), probably because their composition depends on the specific needs or requests of researchers and the screening procedure used by the panel providers (Brandon, Long, Loraas, Mueller-Phillips, & Vansant, 2013). For instance, researchers can request samples from specific countries or regions or select panel participants based on age, gender, or education level.

In order to examine the composition of convenience samples further, I use original data obtained from nine independent samples.<sup>1</sup> More precisely, the data come from two MTurk samples with U.S. participants ( $N_s = 510$  and  $488$ ), four samples from Qualtrics online panels ( $N = 105$  from Switzerland,  $N = 104$  from Germany,  $N = 104$  from Greece, and  $N = 102$  from Spain), and three samples with business students ( $N = 105$  from Canada,  $N = 67$  from Germany, and  $N = 57$  from Switzerland). Furthermore, I compare this original data with data from 16 samples from published articles using the

<sup>1</sup> These data have been collected as part of three large research projects on applicants' experiences during the selection process. All participants were required to have taken part in a selection process in the last year. None of the studies involved any experimental manipulation that could have influenced the scores. I present here only the sample composition and the CWs data, but the full results obtained in those projects (i.e., with other key variables) are currently prepared as three independent manuscripts.

CWs measure:<sup>2</sup> 11 samples using psychology students and five samples using a general adult population (i.e., approached in public places in large cities and sent a survey questionnaire). The demographic composition of each sample type is presented in Table 1.

Participants in both the MTurk samples and the online panels were older than were those from the original business student samples (Cohen's  $d = 0.75$ , 95% CI [0.6; 0.9], and  $d = 1.25$ , 95% CI [1.07; 1.42], respectively; CI = confidence interval) and the published psychology student samples ( $d = 1.75$ , 95% CI [1.67; 1.83], and  $d = 2.41$ , 95% CI [2.29; 2.52]). They were also much more similar to the general population surveys ( $d = -0.12$ , 95% CI [-0.2; -0.04], and  $d = -0.12$ , 95% CI [-0.23; -0.01]). Gender parity was almost achieved in all three original convenience sample types, which was not the case with the published samples (both students and the general population surveys). About half of the participants possessed university/college-level education both with MTurk (53%) and the online panel participants (47%, although it varied from country to country). Regarding employment status, both MTurk samples and online panel samples included mostly employed (48% and 60%) but also unemployed (31% and 30%) and student (21% and 10%) participants (although it varied from country to country in the online panels). No information about education or employment status was provided for the general population surveys. Finally, the ethnicity of the U.S. samples was examined. The composition of MTurk samples was diverse, albeit slightly overrepresenting Whites and Asians but underrepresenting Blacks and Hispanics as compared with the U.S. workforce (U.S. Bureau of Labor Statistics, 2014). Interestingly, a similar conclusion can be reached for the ethnic composition of U.S. student samples in published research (i.e., overrepresentation of Asians but underrepresentation of Blacks and Hispanics).

Altogether, these results suggest that MTurk and online panels may not perfectly mirror organizational samples, for instance, because they still incorporate unemployed individuals or students. However, they offer access to samples that in some aspects resemble the working population more than student samples (e.g., older and likely more experienced). It is for this reason that I-O psychology journals are less reluctant toward masters of business administration samples than undergraduate student samples (Landers & Behrend, 2015). Given the results shown in this commentary, they may

<sup>2</sup> These articles were identified through PsycINFO and Google Scholar. I only included studies where a scale (vs. just one item) was used to measure CWs and where CWs scores (means, standard deviations) and information about the sample were available. The majority of the samples were from the United States and New Zealand. The full references of these articles are not included here to keep this commentary short, but they can be obtained from the author.

**Table 1** Demographics and Competitive Worldviews Data Across Samples

	Original data			Published data from social psychology	
	Crowdsourcing	Online panels	Business students	Psychology students	General population
<i>Demographics</i>					
<i>K</i>	2	4	3	11	5
<i>N</i>	998	415	223	3,440	1,498
Age: <i>M (SD)</i>	30.43 (10.16)	30.35 (6.09)	23.42 (3.59)	19.99 (3.96)	31.94 (12.79)
Gender (% female)	44	49	49	65	57
College education (%)	53	47	N/A	N/A	—
Employment status					
Employed (%)	48	60	N/A	N/A	—
Unemployed (%)	31	30			
Student (%)	21	10			
Ethnicity (%)					
White (%)	76	N/A	N/A	61	—
Asian (%)	10			26	
Black (%)	8			8	
Hispanic (%)	5			4	
Native (%)	1			1	
<i>Competitive worldviews</i>					
<i>M (SD)</i>	2.26 (.58)	2.39 (.51)	2.32 (.54)	2.18 (.55)	2.34 (.74)
Range	3.10	2.40	2.80	2.83	3.56
Reliability (alpha)	.90	.84	.84	.76	.82
Skewness: <i>M (SD)</i>	.11 (.11)	-.25 (.24)	.16 (.27)	—	—
Skewness: <i>z</i> value	0.98	1.05	0.57		
Kurtosis: <i>M (SD)</i>	-.41 (.22)	-.71 (.47)	-.43 (.54)		
Kurtosis: <i>z</i> value	1.85	1.51	0.79		

*Note.* All competitive worldviews (CWs) scores for the original data are based on the original 20-item CWs with a 1–5 Likert scale; CWs scores from published data are mostly based on a sample of items from the original scale using 1–7 or –4–4 Likert scales (transformed into 1–5 scores for comparison); Ethnicity statistics are based on U.S. samples only. Range for the original samples was the observed range from the data, whereas range for the published sample was estimated based on  $M \pm 3SD$ .

want to consider MTurk and online panels as viable alternatives, for the same reason. Moreover, online samples may include a large proportion of employed and unemployed job seekers, which may be particularly relevant for specific research areas (e.g., selection and recruitment, vocational and career adjustment). For instance, 43% of our MTurk and 62% of our online panel participants reported that they were actively looking for jobs at the time of our data collections.

### Convenience Samples Data Quality

Landers and Behrend's (2015) focal article also discussed the issue of data quality when using convenience samples. For instance, they argue that on-line samples like MTurk may reduce problems with range restriction on the variables of interest, an issue sometimes associated with organizational samples because employees may have been selected based on this variable (e.g., cognitive ability). Earlier examinations of MTurk have highlighted that high-quality data can be obtained even with low compensation (e.g., Behrend, Sharek, Meade, & Wiebe, 2011; Buhrmester et al., 2011; Crump, McDonnell, & Gureckis, 2013). However, those studies have focused their effort on reliability alpha, test–retest reliability, or experimental studies but have overlooked range restriction or data normality. Other online panels (e.g., SurveyMonkey or Qualtrics) have also been described as offering researchers access to data of adequate quality (Brandon et al., 2013), but existing investigations have focused on completion rates or correct responses to manipulation check questions.

I provide some additional evidence about the quality of data obtained with MTurk, online panels (Qualtrics) and student samples by examining mean differences, potential range restriction, scale reliability, and data normality (see Table 1). In order to do so, all participants in the nine samples completed the same 20-item CWs scale. CWs capture people's tendency to perceive the world as a competitive jungle characterized by a ruthless struggle to obtain scarce resources (Duckitt et al., 2002). CWs represent a pertinent measure to assess data quality because the factor used to screen participants in the nine samples (i.e., a recent experience with selection) is theoretically unrelated to CWs, thus offering a "clean" example of the risks versus the benefits associated with convenience samples. Like with sample composition, I also compare the original convenience samples with published studies using data from psychology student samples or general adult population surveys.

CWs scores for MTurk were higher than were psychology students' ( $d = 0.16$ , 95% CI [0.09; 0.23]) but were similar to business students' ( $d = -0.11$ , 95% CI [-0.25; 0.04]) and the general population's scores. ( $d = -0.11$ , 95% CI [-0.19; -0.02]). CWs scores for the online panels were higher than were psychology students' ( $d = 0.4$ , 95% CI [0.3; 0.5]) but were similar to business students' ( $d = 0.13$ , 95% CI [-0.03; 0.3]) and the general population's ( $d = 0.08$ , 95% CI [-0.04; 0.18]) scores. MTurk CWs scores were lower than were the online panels' ( $d = -0.24$ , 95% CI [-0.35; -0.12]) score. The theoretical range of CWs scores for the general population was 3.56. Among the various convenience samples, MTurk offered the least range restriction (3.10). Psychology students (2.83), business students (2.80), and online panel data (2.40) offered more range restricted CWs data. The

reliability coefficients obtained with all three types of convenience samples for the CWs measure were excellent and equivalent (if not superior) to coefficients of published studies using psychology students or survey data from the general population. Finally, data normality was examined with skewness and kurtosis measures. In all three sample types (even with rather large samples), converted  $z$  scores were below the traditional significance threshold (i.e., 1.96), suggesting that samples distribution did not deviate from normality.

Altogether, the scores of all three original convenience samples (i.e., MTurk, online panels, and business students) were similar to those of the general adult population, while psychology student samples obtained lower scores. The issue of range restriction was limited for MTurk, but it appeared to be more important for online panels. Moreover, data obtained with all three sample types (including MTurk and online panels) achieved good reliability. Finally, convenience sample data normality was not an issue. Overall, these results offer support to Landers and Behrend's (2015) claims and complement earlier evidence for MTurk (Buhrmester et al., 2011) and online panels (Brandon et al., 2013) data quality. Yet, they are based on only one measure and a small number of samples and should be extended to other measures and samples.

### Conclusion

This commentary aimed at better understanding the potential value and risks associated with convenience samples. Crowdsourcing (e.g., MTurk) represents a fast and inexpensive sampling method, allowing researchers to collect data from quite large samples in only hours for as cheap as a few cents or dollars per participant (Behrend et al., 2011). Moreover, screening questions can also allow researchers to target participants who possess specific characteristics (e.g., a recent experience with selection in the samples presented here). I present here some additional evidence showing that MTurk can offer access to reliable and high-quality data from participants whose demographics resemble the general labor force more than student samples (e.g., older, more diverse). Online panels like Qualtrics represent another alternative, as they offer access to reliable data that also resemble the general labor force. Moreover, although data collection is likely to be more costly with commercial online panels than with MTurk, such sampling strategies may help researchers to access samples from regions that are currently underrepresented in MTurk (e.g., some European countries).

More research is certainly necessary to better understand how and when crowdsourcing and online panels constitute viable alternatives to organizational samples for I-O psychology research. Yet, evidence is (slowly) accumulating to support Landers and Behrend's (2015) call for researchers, reviewers, and editors to not automatically condemn such sampling strategies.

## References

- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43, 800–813. doi:[10.3758/s13428-011-0081-0](https://doi.org/10.3758/s13428-011-0081-0)
- Brandon, D. M., Long, J. H., Loraas, T. M., Mueller-Phillips, J., & Vansant, B. (2013). Online instrument delivery and participant recruitment services: Emerging opportunities for behavioral accounting research. *Behavioral Research in Accounting*, 26, 1–23.
- 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](https://doi.org/10.1177/1745691610393980)
- Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, 8, e57410. doi:[10.1371/journal.pone.0057410](https://doi.org/10.1371/journal.pone.0057410)
- Duckitt, J., Wagner, C., Du Plessis, I., & Birum, I. (2002). The psychological bases of ideology and prejudice: Testing a dual process model. *Journal of Personality and Social Psychology*, 83, 75–93. doi:[10.1037/0022-3514.83.1.75](https://doi.org/10.1037/0022-3514.83.1.75)
- Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 8.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419. doi:[10.2139/ssrn.1626226](https://doi.org/10.2139/ssrn.1626226)
- U.S. Bureau of Labor Statistics. (2014). *Labor force characteristics by race and ethnicity, 2013*. Retrieved from <http://www.bls.gov/cps/cpsrace2013.pdf>

## Inattentive Responding in MTurk and Other Online Samples

Avi Fleischer

*Tetric LLC, Chicago, Illinois*

Alan D. Mead

*Talent Algorithms Inc., Chicago, Illinois*

Jialin Huang

*Illinois Institute of Technology*

The focal article by Landers and Behrend (2015) makes the case that samples collected on microtask websites like Amazon's Mechanical Turk (MTurk) are inherently no better or worse than traditional samples of convenience from university students or organizations. We wholeheartedly agree. However, having successfully used MTurk and other online sources for data col-

Avi Fleischer, Tetric LLC, Chicago, Illinois; Alan D. Mead, Talent Algorithms Inc., Chicago, Illinois; Jialin Huang, Department of Psychology, Illinois Institute of Technology.

Correspondence concerning this article should be addressed to Avi Fleischer, Tetric LLC, Chicago, IL. E-mail: [fleischer@tetric.com](mailto:fleischer@tetric.com)