Examining Validity of MTurk Workers Responses Based on Monetary Reward

Margret Murphy, David Condon,

University of Oregon, Department of Psychology

Introduction

In recent years, an increasing proportion of social science research has been conducted online, and this is particularly true for survey-based research. Recruitment for this research often draws from online crowdsourcing sites like Amazon's Mechanical Turk (Dupuis, Endicott-Popovsky, & Crossler, 2013; Goodman & Paolacci, 2017). Those who respond to these surveys – "MTurkers" – are compensated for their participation. Compensation tends to be quite low, with average hourly pay below one-third of national minimum wage standards in the U.S. (Hara et al., 2017). Variability is also substantial, with allotments ranging as low as $0.01 USD (Semuels, 2018). To date, there has been little research that considers the relationship between compensation rates and data quality (Chmielewski & Kucker, 2019). We sought to evaluate the effect compensation has on the validity of MTurkers' responses to psychological surveys about personality and mental health.

Research Question

Does the amount of compensation paid to Amazon's Mechanical Turk workers affect the quality of data collected from psychological surveys?

Methods

The study design called for data collection from three groups of participants. All three were to be recruited using the same description of our "Human Intelligence Task" (known as a HIT on the MTurk platform), except that each group was to be recruited using differing amounts of payment relative to the U.S. federal minimum wage. The first group was to be paid at an hourly rate equivalent to minimum wage ($7.25/hr), the second at a rate equal to 25% more than minimum wage, and the third at a rate equal to 25% less than minimum wage with an unannounced bonus after the work was completed to bring their total payment up to minimum wage.

Measures: These data were collected as part of a larger project to develop normative values for a measure of personality. As such, participants were administered the 81-item measure of the SAPA Personality Inventory (Condon, 2018), the 10-item PROMIS Global Health measure (Hayes et al., 2009), and the 17-item Comprehensive Health Survey (Goldberg, 2018).

Participants: Data for the first two sub-samples were collected as planned. All of the participants in both groups were residents of the U.S. and self-reported as fluent (99.3%) or nearly fluent (0.7%) in English, as these variables were used to screen participants at the outset of data collection. Together, the two groups included 1,158 participants from 46 states plus the District of Columbia. The minimum wage group (dba MinWage) contained 579 participants (50.9% female), ranging in age from 18 to 78 years (*M* = 45.3, *Mdn* = 47, *SD* = 16.2). The group paid 25% above minimum wage (dba MinWage125) contained 579 participants (54.9% female), ranging in age from 20 to 77 years (*M* = 45.4, *Mdn* = 46, *SD* = 16.3). Though similarly diverse in educational background (ranging from "less than 12 years" to "graduate or professional degree"), MinWage125 had a larger proportion of participants with a college degree or more education (55.4% vs 41.4%).

The 3rd sub-sample was not collected due to the COVID-19 pandemic.

Results

To test the effect of compensation on data quality for psychological surveys, we compared the two samples on several aspects.

Inattentive responding: The survey included one item in the middle that was written to test whether participants were paying attention. The majority of participants answered this item correctly in both samples, though a slightly higher proportion missed it in the sample receiving lower compensation (97.7% vs 99.0%). This difference was not statistically significant (*p* = .08).

Consistency of responding on similar items: Given the use of highly similar items in each scale of the personality measures, we evaluated differences in response patterns by comparing the standardized alphas across samples for each of 27 personality scales. Based on chance, we expected that 1 or 2 of these scales would differ; we found differences (*p* < .05) in 4 scales.

Figure 1

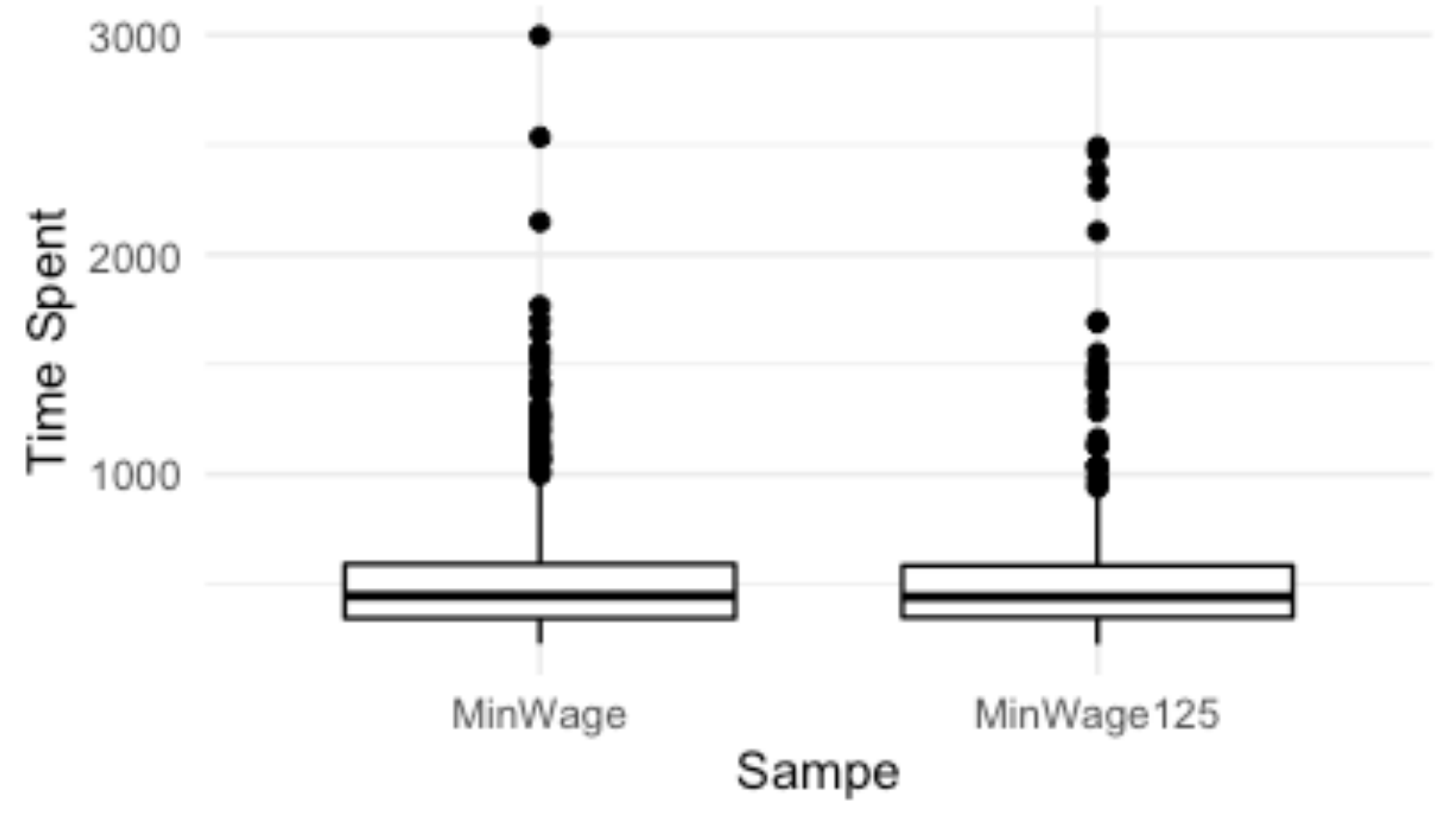
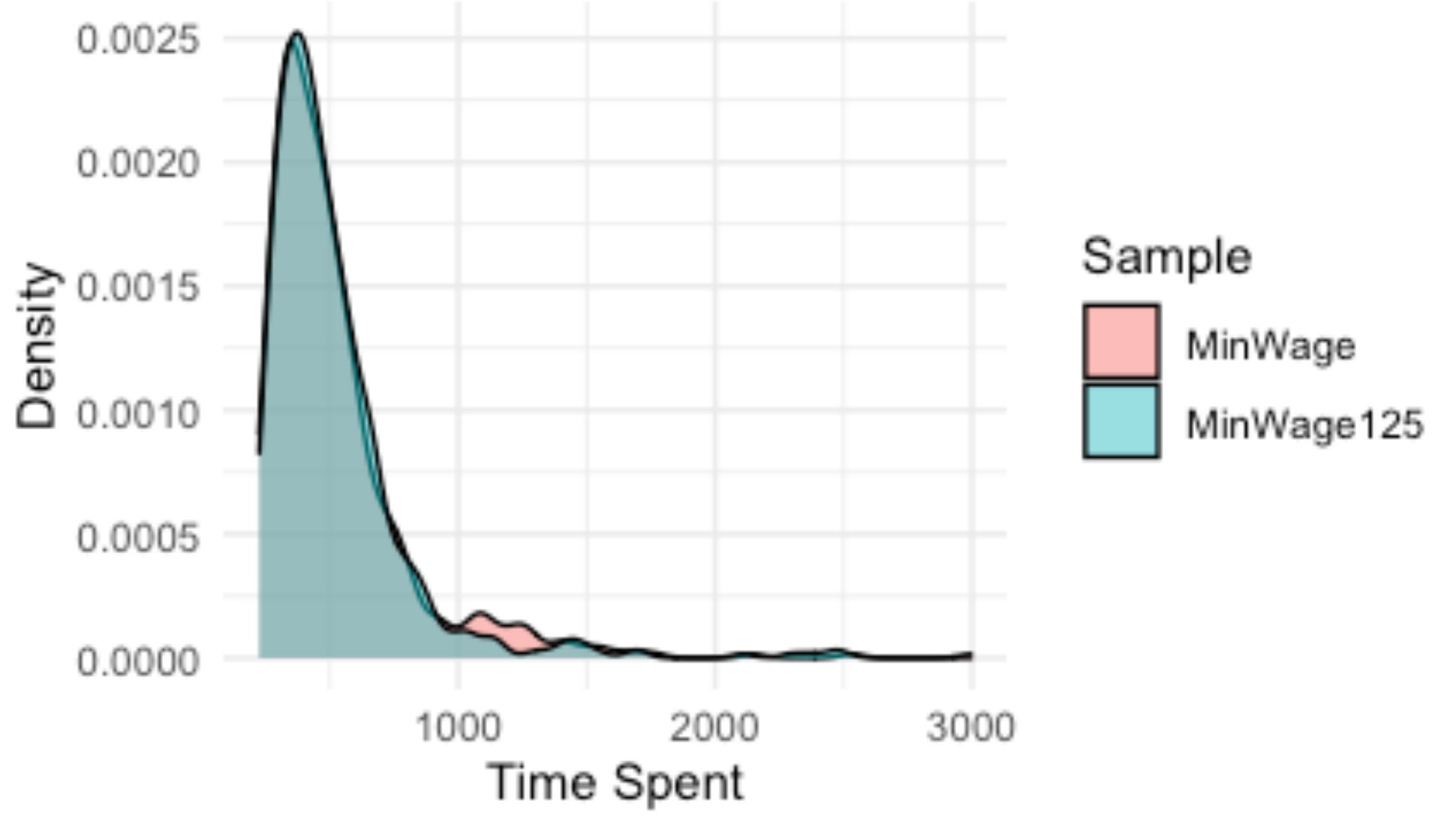


Figure 2



Time Spent on Task: The total time spent on the survey varied considerably within and across the samples. The minimum and maximum times ranged from 1 min 6 secs to 34 hrs 8 min 10 secs in the MinWage sample and from 1 min 11 secs to 56 min 42 secs in the MinWage125 sample. After removing outliers who took an average of more than 30 seconds per item (1 participant in each sample) or less than 2 seconds per item (MinWage: 30 participants, MinWage125: 24 participants), the time spent on task was about 8 min 40 secs in both samples. A t-test comparing means was not significant (*p* = .53). A test for differences between the distributions of the two samples was also not significant (two-sample Kolmogorov-Smirnov test: *p* = .90). Figure 1 shows a boxplot of the time spent by participants in each sample. Figure 2 shows the highly overlapping density distributions of time spent for each sample.

Figure 3

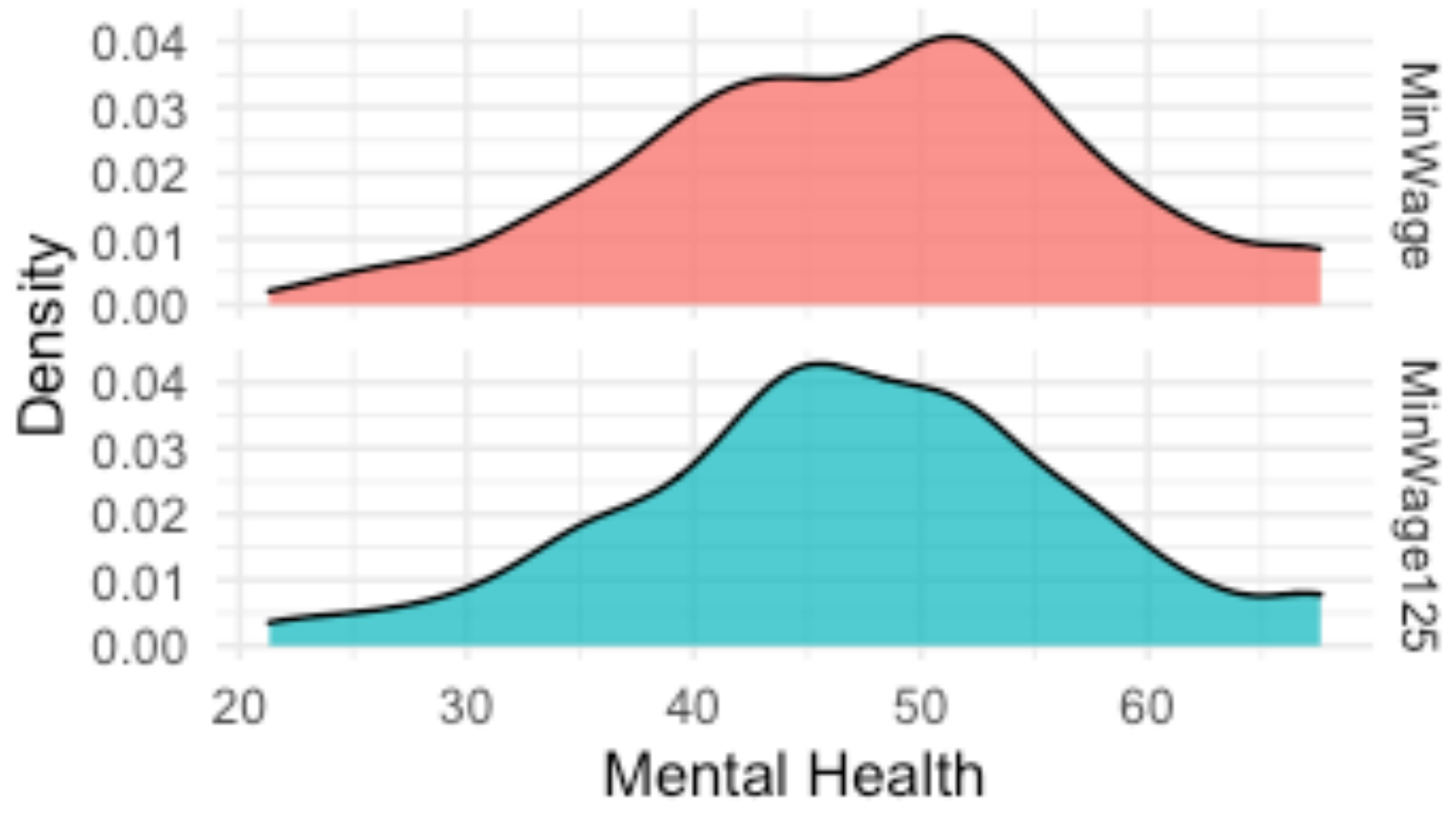
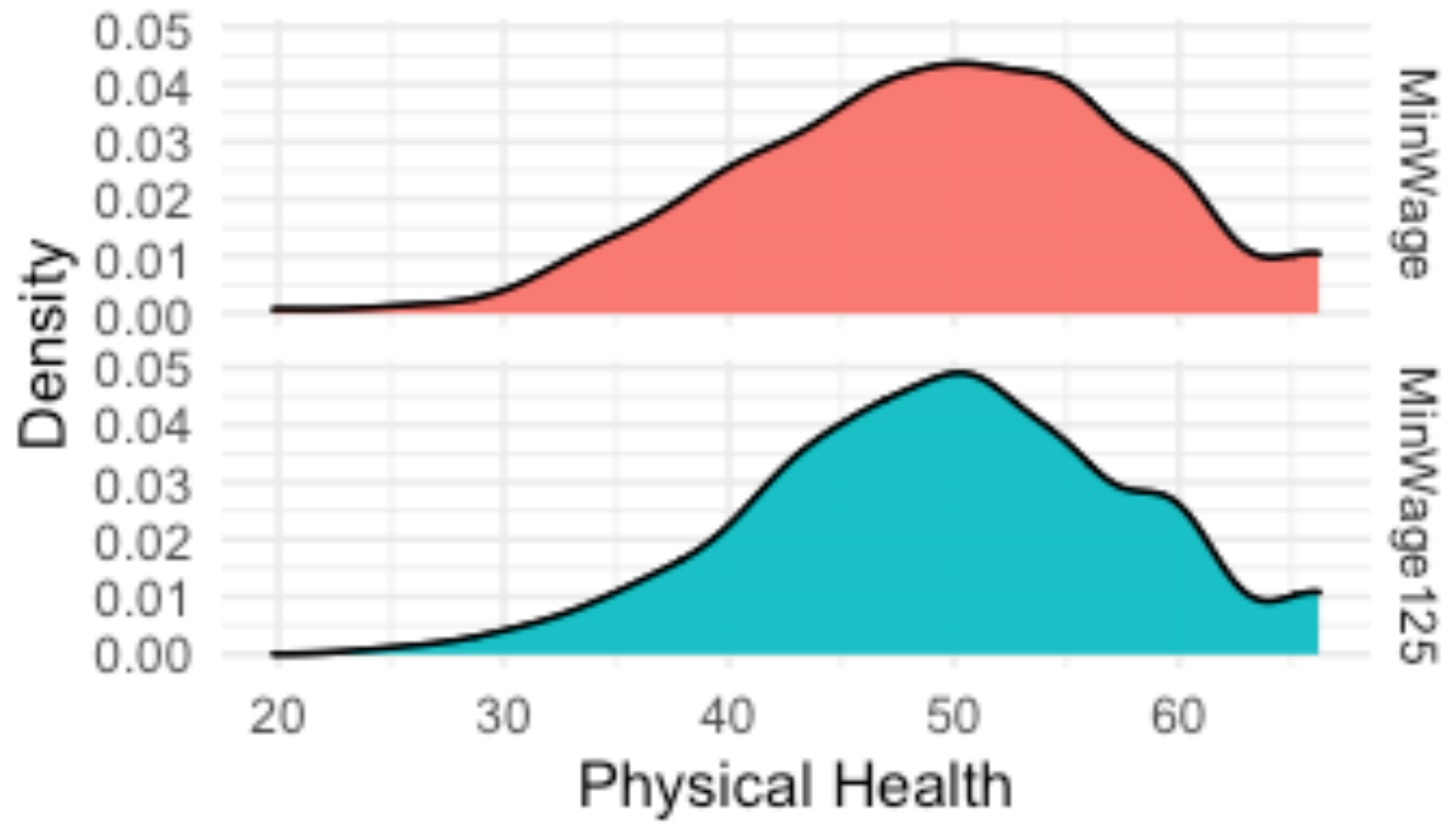


Figure 4



Representativeness in terms of physical and mental health: Participants in both samples completed self-report measures of mental and physical health that have been normed in very large nationally representative cohorts. Using the same tests described above, no statistical differences between samples were found. Figures 3 and 4 show the highly overlapping distributions for Mental and Physical Health respectively. Note that the mean Mental Health score in both samples was approximately 1/3rd SD below the national norm (worse), as can be seen in Fig 3.

Conclusion

While there was a slight difference in attention responding and consistency of responses, the validity of patient responses has a slight inference on compensation rates, though it is not statistically significant enough to state that compensation affects the quality of the data collected. It is with these results that conclude compensation does not impact the validity of participant responses. This research will benefit the discussion of crowdsourcing sites in terms of their validity of retrieving quality data. Future research would benefit this discussion further, suggesting the re-creation of this study across differing crowdsourcing sites.

References & Acknowledgements

References:

Chmielewski, M., & Kucker, S. (2019). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. Social Psychological and Personality Science, 11(4), 464-473.
Condon, D. M. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. PsyArXiv.
Dupuis, M., Endicott-Popovsky, B., & Crossler, R. (2013). An analysis of the use of Amazon's Mechanical Turk for survey research in the cloud. Proceedings of the International Conference on Cloud Security Management: ICCSM (p. 10).
Goldberg, Lewis R., 2018, "CHS.pdf", (17) Comprehensive Health Survey (CHS), https://doi.org/10.7910/DVN/FFIH05/3LWPCC, Harvard Dataverse, V1
Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. Journal of Consumer Research, 44(1), 196-210.
Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-14).
Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Quality of Life Research, 18(7)
Semuels, A. (2018). The internet is enabling a new kind of poorly paid hell. The Atlantic.
Watkins, D. C. (2012). Qualitative Research: The Importance of Conducting Research That Doesn't "Count." Health Promotion Practice, 13(2), 153–158.

Acknowledgements:

I would like to thank David Condon and the PIElab for making me a part of their amazing team! I would also like to acknowledge the symposium for hosting me and allowing me to present this research!