The Journal of Open Source Software

excluder: An R package that checks for exclusion criteria in online data

Jeffrey R. Stevens[1]

1 Department of Psychology, Center for Brain, Biology & Behavior, University of Nebraska-Lincoln

DOI: 10.21105/joss.03893

Software

- Review
- Repository
- Archive

Editor: Daniel S. Katz

Reviewers:

- @danielskatz

Submitted: 04 November 2021
Published: 05 November 2021

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

Collecting survey data online can result in low-quality data. Survey participants may not complete the survey, may complete the survey too quickly or slowly, may not reside in the country they claim, or may use unacceptable screen types. Also, online surveys are plagued by automated bots attempting to complete the surveys while offering worthless data. Researchers collecting online data may want to check their data for these and other potential criteria to exclude problematic data entries. The `excluder` package uses three main function types to mark, check, and exclude data based on seven different exclusion criteria.

Statement of need

Researchers who conduct online surveys may use Qualtrics or other online systems to collect data from participants. Those participants may be recruited directly via listservs or through third party vendors that connect researchers and participants, such as Amazon Mechanical Turk and Prolific. Ensuring good data quality from these participants can be tricky (Aruguete et al., 2019; Chmielewski & Kucker, 2020; Eyal et al., 2021; Gupta et al., 2021). For instance, while Mechanical Turk in theory screens workers based on location (e.g., if you want to restrict your participant pool to workers in the United States), this is not necessarily represented in the data when participant IP addresses are recorded. Also, automated bots are constantly trying to complete online surveys with worthless data. Therefore, researchers may want to screen their data for certain exclusion criteria.

Finding the tools to screen for IP address location can be difficult, and the `excluder` package simplifies working with exclusion criteria based on data that Qualtrics reports, including geolocation, IP address, duplicate records from the same location, participant screen resolution, participant progress through the survey, and survey completion duration. `excluder` is an R (Team, 2021) package based on the `tidyverse` (Wickham et al., 2019) framework that use three primary functions to (1) mark existing files with new columns that flag data rows meeting exclusion criteria, (2) view the subset of data rows that meet exclusion criteria, and (3) exclude data rows that meet exclusion criteria from the data. In addition, `excluder` helps prepare Qualtrics data for analysis and can deidentify the data by removing columns with potentially identifiable information. Though the functionality focuses on data collected by Qualtrics and imported by the `qualtRics` (Ginn & Silge, 2021) package, it is flexible enough for researchers using any source of online survey data.

Acknowledgments

I thank Francine Goh and Billy Lim for comments on an early version of the package, as well as the insightful feedback from rOpenSci staff Mauro Lepore, Joseph O'Brien, and Julia Silge.

This work was funded by US National Science Foundation grant NSF-1658837.

References

Aruguete, M. S., Huynh, H., Browne, B. L., Jurs, B., Flint, E., & McCutcheon, L. E. (2019). How serious is the 'carelessness' problem on Mechanical Turk? *International Journal of Social Research Methodology*, *22*(5), 441–449. https://doi.org/10.1080/13645579.2018.1563966

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01694-3

Ginn, J., & Silge, J. (2021). qualtRics: Download 'Qualtrics' Survey Data. In *GitHub repository*. GitHub. https://github.com/ropensci/qualtRics/

Gupta, N., Rigotti, L., & Wilson, A. (2021). The experimenters' dilemma: Inferential preferences over populations. *arXiv*, 2107.05064. http://arxiv.org/abs/2107.05064

Team, R. C. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686