The Journal of Open Source Software

trustmebro: An R package for Plausibility Checks and Cleaning of Subject-Generated ID Codes

Annemarie Pläschke [1*] **and Tobias Brändle** [1*]

1 Trusted Third Party (Vertrauenstelle nach §98a HmbSG), Authority for Schools, Family and Vocational Education in Hamburg, Germany * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 30 September 2025
Published: unpublished

Summary

`trustmebro` is an R package designed to simplify data wrangling tasks involving subject-generated ID codes (SGICs) and other identifiers. It provides functions for plausibility checks of SGICs, helping to ensure the reliability and consistency of identifier data. Alongside validation tools, trustmebro offers utilities for streamlined recoding, facilitating the management of messy identifier-related data in research workflows.

Statement of need

Subject-generated ID codes (SGICs) are widely used in research contexts where individual-level data from different sources must be linked without exposing personally identifiable information (Calatrava et al., 2022; Carifio & Biron, 1978, 1981; Kearney et al., 1984; Schnell et al., 2010). However, SGICs are often prone to errors, inconsistencies, and formatting issues, which can compromise data quality and reduce the validity of downstream analyses (Audette et al., 2020; Dilorio et al., 2000; Kristjansson et al., 2014; Little et al., 2022; Yurek et al., 2008).

`trustmebro` was developed to address these challenges in the context of educational data linkage at the Trusted Third Party (Vertrauensstelle) of the Authority for Schools and Vocational Education in Hamburg, Germany (Fickermann & Maritzen, 2014). By providing functions for inspecting and validating SGICs, as well as tools for handling common identifier-related tasks, the package helps ensure that data integration processes are both reliable and efficient. While originally created for educational research use cases, `trustmebro` can be applied broadly in any setting where SGICs or similar identifiers play a central role in data wrangling.

Acknowledgements

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

References

Audette, L. M., Hammond, M. S., & Rochester, N. K. (2020). Methodological Issues With Coding Participants in Anonymous Psychological Longitudinal Studies. *Educational and Psychological Measurement*, *80*(1), 163–185. https://doi.org/10.1177/0013164419843576

Calatrava, M., De Irala, J., Osorio, A., Benítez, E., & Lopez-del Burgo, C. (2022). Matched and Fully Private? A New Self-Generated Identification Code for School-Based Cohort

Studies to Increase Perceived Anonymity. *Educational and Psychological Measurement*, *82*(3), 465–481. https://doi.org/10.1177/00131644211035436

Carifio, J., & Biron, R. (1978). Collecting Sensitive Data Anonymously: The CDRGP Technique. *Journal of Alcohol and Drug Education*, *23*(2), 47–66.

Carifio, J., & Biron, R. (1981). Collecting Sensitive Data Anonymously: Further Findings on the CDRGP Technique. *Journal of Alcohol and Drug Education*, *25*(2), 38–70.

Dilorio, C., Soet, J. E., Van Marter, D., Woodring, T. M., & Dudley, W. N. (2000). An evaluation of a self-generated identification code. *Research in Nursing & Health*, *23*(2), 167–174. https://doi.org/10.1002/(SICI)1098-240X(200004)23:2%3C167::AID-NUR9%3E3.0.CO;2-K

Fickermann, D., & Maritzen, N. (Eds.). (2014). *Grundlagen für eine daten- und theoriegestützte Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ)*. Waxmann. ISBN: 978-3-8309-3053-2

Kearney, K. A., Hopkins, R. H., Mauss, A. L., & Weisheit, R. A. (1984). Self-Generated Identification Codes for Anonymous Collection of Longitudinal Questionnaire Data. *Public Opinion Quarterly*, *48*(1B), 370–378. https://doi.org/10.1093/poq/48.1B.370

Kristjansson, A. L., Sigfusdottir, I. D., Sigfusson, J., & Allegrante, J. P. (2014). Self-Generated Identification Codes in Longitudinal Prevention Research with Adolescents: A Pilot Study of Matched and Unmatched Subjects. *Prevention Science*, *15*(2), 205–212. https://doi.org/10.1007/s11121-013-0372-z

Little, M. A., Bonilla, G., McMurry, T., Pebley, K., Klesges, R. C., & Talcott, G. W. (2022). The Feasibility of Using Self-Generated Identification Codes in Longitudinal Research With Military Personnel. *Evaluation & the Health Professions*, *45*(4), 354–361. https://doi.org/10.1177/01632787211031625

Schnell, R., Bachteler, T., & Reiher, J. (2010). Improving the Use of Self-Generated Identification Codes. *Evaluation Review*, *34*(5), 391–418. https://doi.org/10.1177/0193841X10387576

Yurek, L. A., Vasey, J., & Sullivan Havens, D. (2008). The Use of Self-Generated Identification Codes in Longitudinal Research. *Evaluation Review*, *32*(5), 435–452. https://doi.org/10.1177/0193841X08316676