# EdSurvey: an R package to analyze Large-scale educational assessments data

Paul Bailey [1][¶], **Charles Blankenship** [1], **Eric Buehler** [2], **Ahmad Emad** [3], **Tom Fink** [1], **Huade Huo** [1], **Claire Kelley** [4], **Michael Lee** [5], **Sun-joo Lee** [6], **Yuqi Liao** [1], **Alex Lishinski** [7], **Trang Nguyen** [8], **Emmanuel Sikali** [9], **Blue Webb** [1], **Qingshu Xie** [1], **Sinan Yavuz** [1], **Jiao Yu** [10], and **Ting Zhang** [1]

1 American Institutes for Research **2** Memorial Sloan Kettering Cancer Center **3** Google LLC **4** Child Trends **5** Independent researcher **6** United Nations Development Programme **7** University of Tennessee Knoxville **8** University of Massachusetts Amherst **9** National Center for Education Statistics, USA **10** JPMorgan Chase & Co. **¶** Corresponding author

Summary

Data from large-scale educational assessment programs, such as the National Assessment of Educational Progress (NAEP) and Progress in International Reading Literacy Study (PIRLS), require special methods to conduct statistical analyses because of their scope and complexity. The EdSurvey package gives users functions to perform analyses that account for both the complex sample survey design and the use of plausible values. The EdSurvey package also seamlessly takes advantage of the LaF (van der Laan, 2020) package to read in data only when required for an analysis. Users with computers that have insufficient memory to read in entire datasets can still do analyses without having to write special code to read in just the appropriate variables. This is all addressed directly in the EdSurvey package — behind the scenes and without any special tuning by the user.

Statement of need

There is a significant demand for accessible and user-friendly software to process and analyze large-scale educational assessment data. These assessments provide valuable insights for policy-makers, education practitioners, and researchers to understand and explore various education-related issues throughout childhood to adulthood, in schools, homes, and neighborhoods.

To access these data with ease, perform data analysis, and test their hypotheses, various free tools have been supplied by assessment institutes, or government agencies, including the International Database IDB Analyzer, NCES Electronic Code Books, NAEP Data Explorer, International Data Explorer, WesVar, and the AM software. Most of these free tools are menu-driven software that do not create a shareable code. They were created with the average statistical user in mind and are limited in the types of the statistical analyses that they can support. There are also commercially available software packages (e.g., SPSS, SAS, or STATA) for data merging, cleaning, and manipulation but many analysts use them for data manipulation only and then choose to import the data into tools such as WesVar or AM to account for complex sample designs and use the plausible values.

The EdSurvey package was conceived and developed to group in one place all the methodologies and techniques that a researcher needs to access, process, manipulate, and analyze these educational databases. EdSurvey can handle most complex samples, plausible value estimations,

and longitudinal data methodologies. EdSurvey runs in R, a programming language licensed under the GNU General Public License, and is widely used by academia and the research communities. The package is tailored to the processing and analysis of large-scale assessment data with appropriate procedures. It is built as a one-stop shop for downloading, processing, manipulation and analysis of survey data. Other packages in R will analyze large-scale assessment data, including survey (Lumley, 2004), lavaan.survey (Oberski, 2014), svyPVpack (Reif & Peterbauer, 2014), BIFIE (BIFIE et al., 2019), intsvy (Caro & Biecek, 2017) and RALSA (Mirazchiyski & INERI, 2021). Among these packages, some have limited data coverage (e.g., tailored to international large-scale assessments only), and some require user input about the survey design and plausible values before they can be used. EdSurvey was developed to analyze all the large-scale assessments that the United States participates in under NCES, and it incorporates a complex sampling design and plausible value methodology seamlessly. The advantages of EdSurvey include:

- allowing for data manipulation inside and outside the package;
- minimizing memory footprint by only reading in required data;
- enabling users to search the names and labels of variables, view frequencies and percentages of response categories of variables, and visualize the data;
- performing complex sample analysis operations;
- computing analyses with plausible values;
- performing multilevel analyses and modeling;
- expanding functions and data supports to meet the needs of analysts with varying levels of expertise; and
- incorporating sampling weights and performing recommended standard error estimations via replicate weighting or the Taylor Series method.

Workflow

Recognizing that researchers using R statistical software come with varying levels of experience, the EdSurvey package has provided multiple workflows to aid in this process of conducting survey analysis. The following graphic details the recommended workflows with corresponding EdSurvey functions.
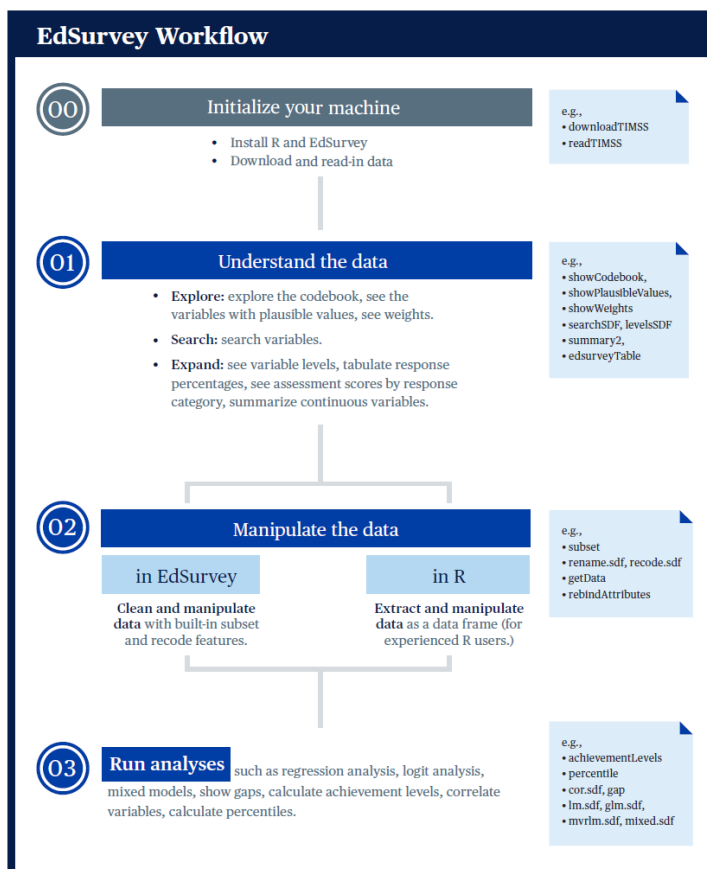
EdSurvey Workflow

00 Initialize your machine
- Install R and EdSurvey
- Download and read-in data

e.g.,
- downloadTIMSS
- readTIMSS

01 Understand the data
- **Explore:** explore the codebook, see the variables with plausible values, see weights.
- **Search:** search variables.
- **Expand:** see variable levels, tabulate response percentages, see assessment scores by response category, summarize continuous variables.

e.g.,
- showCodebook,
- showPlausibleValues,
- showWeights
- searchSDF, levelsSDF
- summary2,
- edsurveyTable

02 Manipulate the data

in EdSurvey	in R
Clean and manipulate data with built-in subset and recode features.	**Extract and manipulate data** as a data frame (for experienced R users.)

e.g.,
- subset
- rename.sdf, recode.sdf
- getData
- rebindAttributes

03 **Run analyses** such as regression analysis, logit analysis, mixed models, show gaps, calculate achievement levels, correlate variables, calculate percentiles.

e.g.,
- achievementLevels
- percentile
- cor.sdf, gap
- lm.sdf, glm.sdf,
- mvrlm.sdf, mixed.sdf

Figure 1: EdSurvey workflow

The workflow has three components:

1. Understanding the data
2. Preparing the data for analysis
3. Running the analysis

Note that when preparing data for analysis, the EdSurvey package provides functions for users to manipulate their data. Experienced R programmers might prefer to extract and manipulate their data using other R methods or supplementary packages. Each method is supported.

Conclusion

The EdSurvey package provides accessible and user-friendly functions to analyze large-scale educational assessment data, accounting for the complex sample survey design and the use of plausible values. For detailed function walkthroughs and the education studies supported by EdSurvey, consult the EdSurvey User's Guide (Lee et al., 2022).

Acknowledgements

References

BIFIE, Robitzsch, A., & Oberwimmer, K. (2019). *BIFIEsurvey: Tools for survey statistics in educational assessment*. https://CRAN.R-project.org/package=BIFIEsurvey

Caro, D. H., & Biecek, P. (2017). Intsvy: An r package for analyzing international large-scale assessment data. *Journal of Statistical Software*, *81*(7), 1–44. https://doi.org/10.18637/jss.v081.i07

Lee, M., Zhang, T., Bailey, P., Buehler, E., Fink, T., Huo, H., Lee, S., & Liao, Y. (2022). *EdSurvey user's guide*. https://naep-research.airprojects.org/Portals/0/EdSurvey_A_Users_Guide/_book/index.html

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1–19.

Mirazchiyski, P. V., & INERI. (2021). *RALSA: R analyzer for large-scale assessments*. https://cran.r-project.org/web/packages/RALSA/index.html

Oberski, D. (2014). lavaan.survey: An R package for complex survey analysis of structural equation models. *Journal of Statistical Software*, *57*(1), 1–27. http://www.jstatsoft.org/v57/i01/

Reif, M., & Peterbauer, J. (2014). *svyPVpack: A package for complex surveys including plausible values*. https://github.com/manuelreif/svyPVpack

van der Laan, J. (2020). *LaF: Fast access to large ASCII files*. https://CRAN.R-project.org/package=LaF