Data Cleaning in Mathematics Education Research: The Overlooked Methodological Step

Aleata Hubbard
WestEd

The results of educational research studies are only as accurate as the data used to produce them. Drawing on experiences conducting large-scale efficacy studies of classroom-based algebra interventions for community college and middle school students, I am developing practice-based data cleaning procedures to support scholars in conducting rigorous research. The poster identifies common sources of data errors in mathematics education research and offers a framework and related data cleaning process designed to address these errors. I seek feedback on the framework and discussion around data cleaning techniques used by other RUME scholars in their research and in the preparation of future researchers.

Key words: Research methodology, Efficacy studies, Algebra

Screening data for potential errors and ensuring anomalies do not influence analyses is an essential step of the research cycle (Wilkinson, 1999). Despite the importance of data cleaning in rigorous research practice, most methodology courses only give cursory attention to the topic (Osborne, 2012). I am developing practice-based data cleaning processes to support scholars in implementing rigorous research in classroom settings. Specifically, I ask: (1) What are the sources of data errors in educational research studies conducted in authentic mathematics learning environments? and (2) How can a data cleaning process be designed to consistently produce accurate, reliable, confidential, and timely datasets?

The framework presented in this poster was informed by two large-scale efficacy studies. Study A was a three-year, nationwide study involving over 10,000 middle school students and 180 mathematics teachers. Study B is a two-year, statewide study of community college elementary algebra courses. During Study A, a list of data related challenges and their associated resolutions was compiled and used to inform the data cleaning process currently used in Study B. Four common sources of data errors appeared in both studies: variations in assessment administration; participant mobility; multiple participant names; and use of external vendor systems. The following data cleaning process was developed to identify and repair these issues:

1) Create visually distinct instrument forms; indicate administration format in final data sets;
2) De-identify study data as early as possible in the data collection process;
3) Compare record counts against participant lists to identify missing and extra records;
4) Check data files for missing values, missing data columns, and extra data columns;
5) Check identifier columns for duplicate values;
6) Transform categorical values into pre-determined standard values;
7) Flag records with errors;
8) Establish a review process so data cleaning work can be checked by another person.

The data cleaning process and taxonomy of common data error sources offered here can provide a framework for other researchers to evaluate their current data management strategies. Furthermore, I hope this work can spark discussion around more comprehensive methodology training for future researchers. I also seek feedback on ways to communicate the process and information on how others in the RUME community handle data cleaning issues in their work.

References

Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data.* SAGE Publications, Inc.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594