Reproducibility in the era of big data: Lessons for developing robust data management and data analysis procedures

Presented to BigSurv18

26 October 2018

Our Team

- Sandra Chafouleas, PhD
  - Co-PI: Neag School of Education, University of Connecticut
- Amy Briesch, PhD
  - Co-PI: Bouvé College of Health Sciences, Northeastern University
- Betsy McCoach, PhD
  - Co-PI: Neag School of Education, University of Connecticut
- Jennifer Dineen, PhD
  - Co-PI: Dept. of Public Policy, University of Connecticut
- Helene Marcy, MPP
  - Project Manager: Neag School of Education, University of Connecticut
- And many other collaborators…

National Exploration of Emotional/Behavioral Detection in School Screening (NEEDs^2)

Funded by The U.S. Department of Education, National Center for Education Research Institute of Education Sciences (R305A140543)

Defining SEB for School-Based Purposes

Summary Rationale for Our Project

Before SEB screeners continue to be developed and evaluated, it is critical that teachers, parents, school administrators and mental health personnel, community stakeholders, researchers, and policy-makers understand ***if and how*** these screeners are being used, and ***what factors influence*** screener usage and student outcomes.

Our Project: Goal 1 (Exploration)

https://needs2.education.uconn.edu/

Arm 1

- RQ1: Nationally, what do state and district-level priorities look like with regard to school-based behavior policy?

Arm 2

- RQ2: Nationally, do school districts incorporate behavior screening practices? If so, what do those practices look like at elementary and secondary levels?
- RQ4: What do key stakeholders perceive as the intended purpose, value, and usability of school-based behavior screening? For those implementing practices, what is the perceived effectiveness?

Arm 3

- RQ3: Does implementation of behavior screening practices predict student behavioral outcomes? If so, do practices serve as a partial mediator and moderator for district characteristics, usability, and behavior curricula practices?

Challenges In The Era Of Big Data

"The era of big data challenges the way we live and interact with the world."

Victor Mayer-Schönberger, Big Data: A Revolution That Will Transform How We Live, Work and Think (2013)

- The volume and variety of Big Data sources contributes to this messiness.
- Big data are typically initially collected for use outside of research and are often unstructured.
- Big Data often possesses differing characteristics than data from traditional survey designs.
- Education: Lack of theory and frameworks regarding how to manage Big Data sources.

The Data

Administrative Datasets

- 2013-2014 National Center for Education Statistics (NCES) Common Core of Data (CCD)
- 2015-2016 Civil Rights Data Collection (CRDC)
- Stanford Education Data Archive (SEDA)
- district-procured special education data.

Survey Datasets

- School District Administrators
- School Building Administrators
- School Support Staff
- Teachers
- Parents

Reliability and Validity

- Reliability refers to consistency or precision of response. Reliability of administrative data is a function of the precision with which it was gathered and documented. Multiple factors influence reliability of administrative data. (E.g.- the coarseness with which continuous data such as percentages are reported influences the reliability of the data.)
- Validity refers to the appropriateness of the inferences made from the data. Validity of administrative data is a function of the degree to which the administrative data match the constructs of interest.
  - are administrative data an adequate proxy for the constructs of interest?
  - Carefully consider the validity of administrative data up front
  - Factors may influence the validity of administrative data:
    - The closeness of the proxy variable to the true variable of interest,
    - The timing of data collection
    - Level of aggregation of the data
    - And more…!

Fallacy 1: *More data are better!*

Too much data creates logistical and analytic issues such as:

- Confusion among the research team (and among future users)
- Wasted resources documenting the provenance of variables never to be used

Recommendations:

- Import only necessary variables into the master dataset
- Avoid having multiple versions of the same variable **or** have a plan for using them in tandem (e.g.- EFA/PCA)
- Spend time in the administrative data set and documentation prior to survey data collection

Fallacy 2: *Merging is about matching by IDs and getting the columns to align.*

Merging Big Data with traditional survey data provides challenges such as:

- Data sources with formats inconsistent with survey dataset(s)
- Data sources with varying levels of quality
- Data sources with inconsistent documentation

Recommendations:

- Clean first, merge later
- Harmonize your data

Fallacy 3: *Saving your syntax is enough to ensure reproducibility.*

Creating a master data file from multiple data files from multiple sources creates the following challenges:

- Many analytic decisions and processes are not syntax based
- Syntax and data dictionaries do not provide a place to store a multi-paragraph description of methodological and substantive decisions

Recommendations:

- Create a separate text document that describes the methodology and can be stored with the data (e.g. Variable Notes or R Markdown)
- Incorporate as much metadata as possible into the data file (e.g.- Stata notes)

Fallacy 4: *Transparency in your process ensures transparency in your final product.*

Transparency in projects that involve both traditional survey research and big data possess additional complexities:

- Data gathered from multiple modes and methodologies provide documentation challenges
- Data found, not made, leaves researchers who aim to be transparent at the mercy of the original data collectors

Recommendations:

Prior to deciding which data to include, researchers need to thoroughly vet administrative data and its documentation to:

- Asses transparency potential
- Ensure measures actually capture the necessary information

Fallacy 5: *Administrative data is higher quality than self-reported data*

Combining Big Data with survey data introduces additional reliability and validity challenges:

- Multiple, additional, factors influence reliability of administrative data (coarseness, aggregation, timeliness)
- Various factors influence the validity of administrative data (closeness, timing, and aggregation)

Recommendations:

- Apply a framework of reliability and validity to administrative data to understand data quality issues prior to survey instrument development.

Fallacy 6: *If there is relevant administrative data it will help answer your research question.*

Using available administrative data that do not clearly map onto their theoretical framework presents analytic challenges:

- Administrative data often exist as a composite variable that cannot be disentangled
- Administrative data often lacks information about quality, dosage, or degree of implementation

Recommendations:

- A comparison of administrative data limitations with the limitations of potential survey items should take place prior to instrument development

Recommendations

Recommendations

- Plan, plan, plan!
- Know your data
- Be thoughtful about merging
- Master data file vs. Analysis files
- Reproducibility!
- Evaluate the quality of outside data– consider both reliability and validity

Questions, Comments, Contact…

https://needs2.education.uconn.edu/

Funded by The U.S. Department of Education, National Center for Education Research Institute of Education Sciences (R305A140543)