Integration of Multiple Data Sources to Inform a Responsive Design

2018 Federal Committee on Statistical Methodology Research and Policy Conference

Peter Siegel, Jennifer Wine
How can data from administrative and commercial records be integrated to help improve survey data quality?

Two National Academies reports in 2017 focused on using multiple data sources for federal statistics:


Once researchers have identified supplemental data sources, and determined how to link the data with sample members and their survey data, the next question is how to use the extensive set of data productively.

Response unit can be defined using all sources.

Multiple sources can be used to inform decision making during data collection.
1. Sample member completing the survey interview under some rule for determining what constitutes ‘complete’
   
   Data collected from other sources may then be used to fill in any missing values prior to imputation.

2. Sample member with sufficient data from any source to be judged complete
   
   Impute to compensate for missing interview data.
Concerns With Defining a Respondent Without an Interview

**Fails to address any bias** among the interview nonrespondents

**Ignores the potential bias** that may be due to missing data from other data sources

**Ethical and privacy issues** because an interview nonrespondent may not know that his/her data are being obtained from administrative sources and has not explicitly given consent to do this
Overview

Conducted by the National Center for Education Statistics (NCES)

Collects data on how students and their families pay for postsecondary education

Two-stage sample selection process

- 2,000 institutions
- 122,000 students
- Students sampled on a flow basis over several months
NPSAS:16 – Data Sources

1. Interview
2. Student records from institutions
3. Central Processing System (CPS)
4. National Student Loan Data System (NSLDS)
5. National Student Clearinghouse (NSC)
6. College Board (SAT)
7. ACT
8. Veterans Benefits Administration (VBA)
NPSAS:16 - Respondents

**INTERVIEW**
- Completed full or abbreviated interview
- Response rate - 66%

**STUDY**
- Use data from any source
- Data for 3 critical variables
- Data for at least 8 of 15 additional variables
- Response rate – 93%
Impute interview for study members with no interview

A lot of imputed data

Conduct item-level nonresponse bias analysis

No multiple imputation for analysis files
Institution and student response rates are primary indicators of data collection success.

Study member nonresponse bias less of a concern with a 93% response rate.

Implemented a two-pronged approach to consider variance reduction.
Prong 1: Abbreviated Interview
Collecting data needed to be a study member

Prong 2: Abbreviated Interview
Collecting data to improve data quality for study members who were interview nonrespondents
Prong 1 – Abbreviated Interview

Motivation – collect some interview data for sample members who were ‘close’ to being study members

Abbreviated interview contained only items needed to qualify as a study member
Prong 1 – Case Targeting

- Tracked accumulation of variables, from each source, needed for study membership
- Interview nonrespondents with at least 6 of the 11 variables needed for study membership were offered abbreviated interview
- Cases were identified at 4 points in time, based on when they began data collection
Prong 2 – Abbreviated Interview

Motivation – collect some interview data for study members who were likely to have high imputation variance

Created a superset of interview-only variables based on analytic importance

Performed multiple imputation on 14 variables available from previous NPSAS data
9 of 14 variables had mean RSEs greater than 25%
Prong 2 – Abbreviated Interview (cont.)

Other important variables new to NPSAS

\[9 + 9 = 18\]

Variables included in abbreviated interview
Prong 2 – Multiple Imputation

On the fly during data collection

Using raw data

For study members who were interview nonrespondents

For 15 of 18 variables (3 had imputation issues for a subset of students)

To estimate the extent of variability present if sample member remained interview nonrespondent
Prong 2 – Case Targeting

Sample members with high variation across 15 imputed items received abbreviated interview

Cases were identified at 3 points in time, based on when they began data collection

Prong 1 cases who completed abbreviated interview were asked to continue with prong 2 abbreviated questionnaire
Conclusions

✓ Results forthcoming, but promising

✓ Important to think through issues that will be raised as data from administrative and commercial sources are used more frequently with survey data

✓ Respondents do not have to be defined based solely on interview data

✓ Responsive design can focus on improving quality of data instead of, or in addition to, reducing nonresponse bias

✓ Best approach when using data from multiple sources will vary based on the study
Questions

Peter Siegel
siegel@rti.org

Jennifer Wine
jennifer@rti.org