Integration of Survey and Administrative Data Collection:

A Responsive and Adaptive Survey Design with Multiple Imputation

Andy Peytchev

Acknowledgment

- Collaborators: Peter Siegel, Jennifer Wine, Jeremy Porter

- The larger NPSAS team

Motivation

Approaches to direct effort and protocol changes during data collection:

- Limited auxiliary information on full sample, typically no Y-variables
  - Focus on information based on respondents (e.g., phase capacity based on observed data)
  - Focus on sample balance on demographic characteristics and interviewer observations, when available (e.g., response propensities, CVs, partial R-indicators)

Motivation continued

- Some surveys incorporate administrative data that include key variables
  - The end data product is a combination of survey and administrative data
  - Administrative data can substitute for some survey variables for nonrespondents, and can be highly correlated with other survey variables
  - Focus should be on what remains unknown for the survey nonrespondents

The 2015-2016 National Postsecondary Student Aid Study (NPSAS:16)

- Heavily dependent on demographic and student aid administrative data: survey estimates are produced from survey respondents and a large proportion of the survey nonrespondents, as long as a minimum amount of administrative data are available for them
  - Survey respondents vs. study members
- Extensive reliance on imputation
  - Over half of the values for some variables
  - Amount of missing data varies across study members
  - Single imputation
  - => Substantial uncertainty in some imputed values that is not reflected in the variance estimates

Concepts in the Proposed Approach

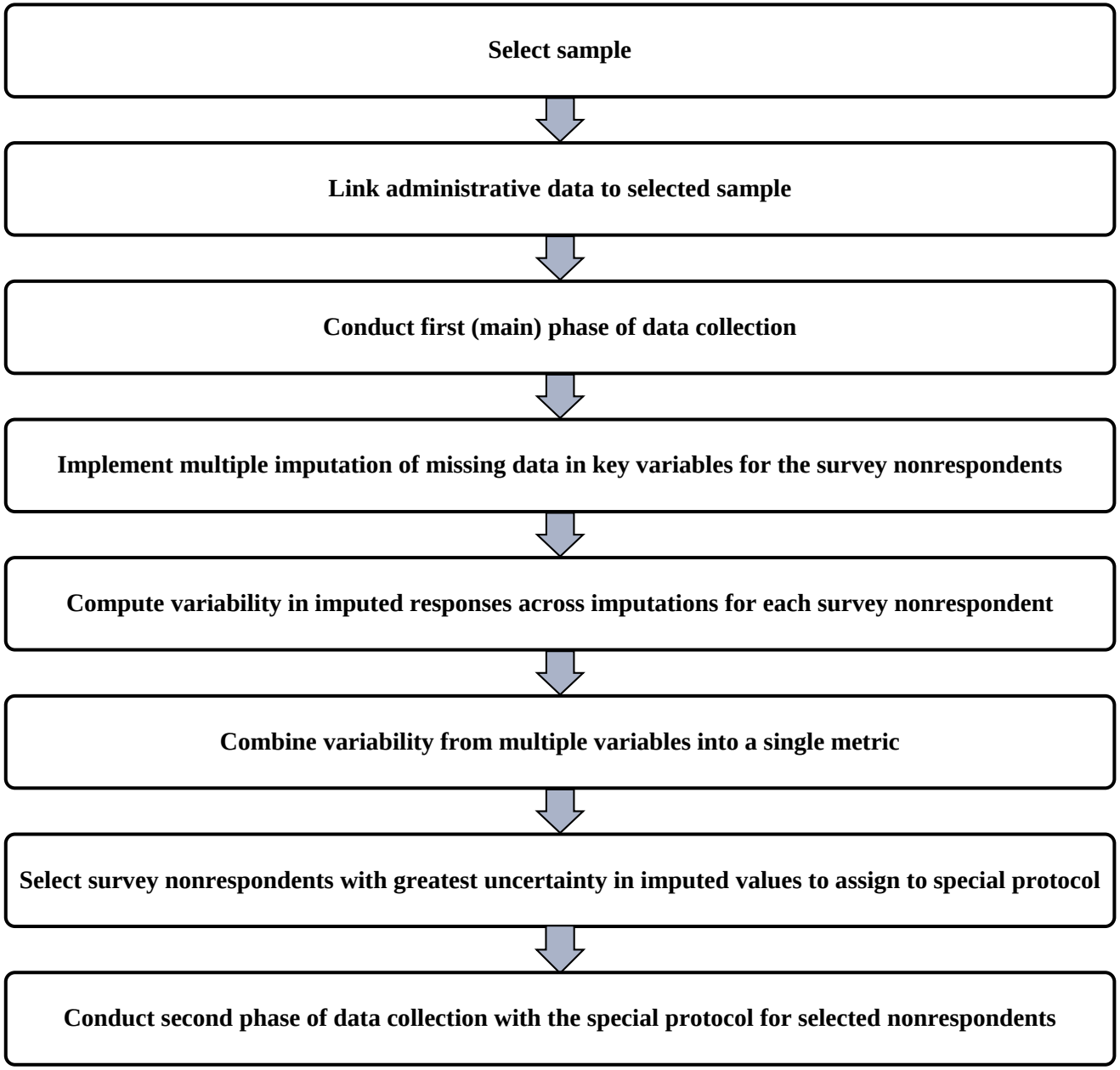
1. Multiple imputation (MI) and Fraction of Missing Information (FMI)
   - Summary measure for uncertainty due to imputation

$$FMI_M = (1 + \frac{1}{M}) \frac{B_M}{T_M}$$

2. Responsive and adaptive survey design (RASD)
   - Additional phase in data collection with a different protocol
   - Target sample cases deemed to be of greatest value to the study

3. MI-based RASD to minimize FMI
   - Compute uncertainty in imputed values at the study member level
   - Target cases with greatest uncertainty with a different protocol

Overview of the Proposed Approach



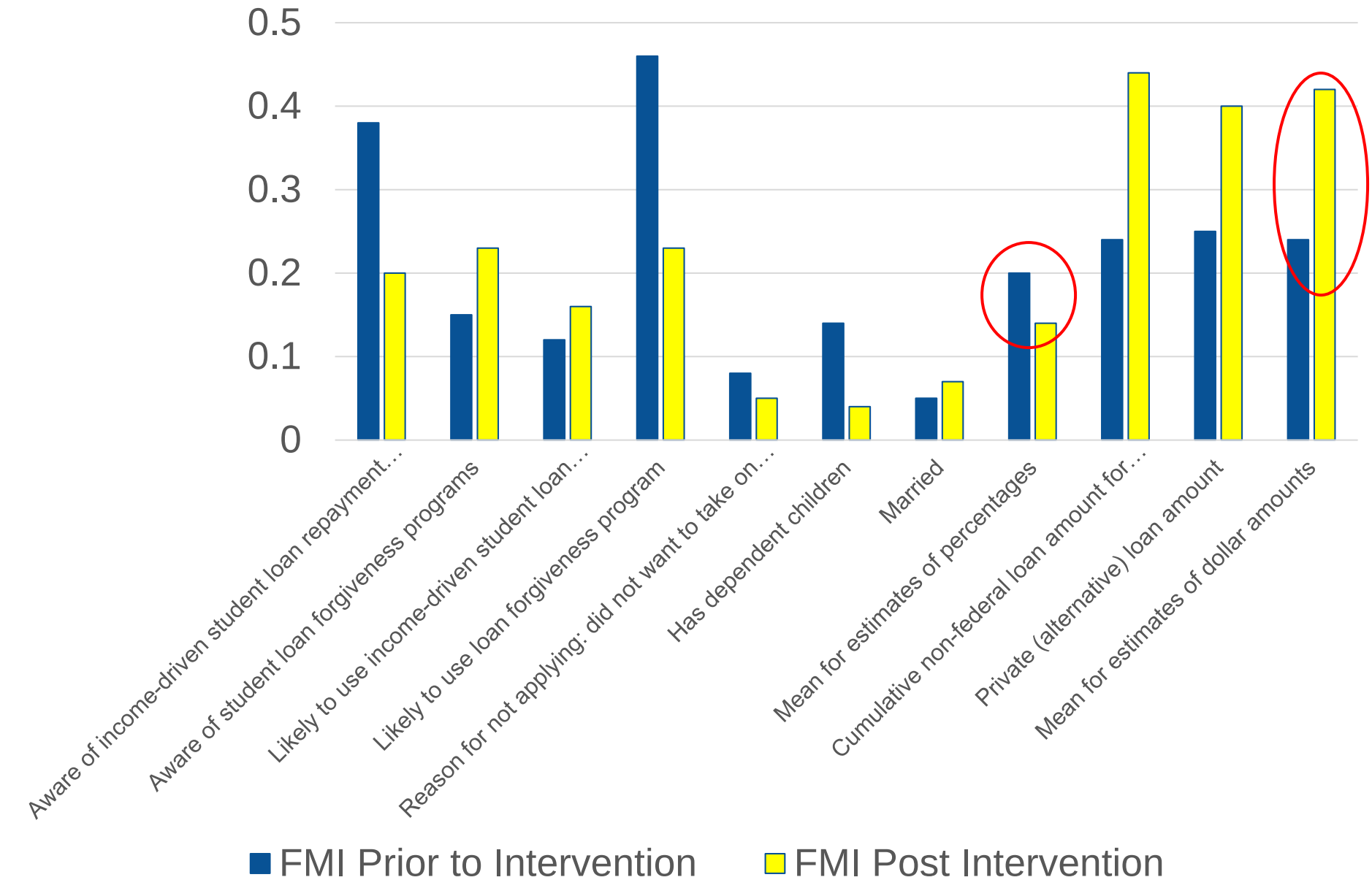
A Few Caveats

- Suboptimal imputation method for this purpose
- Errors in specification of the imputation model during data collection

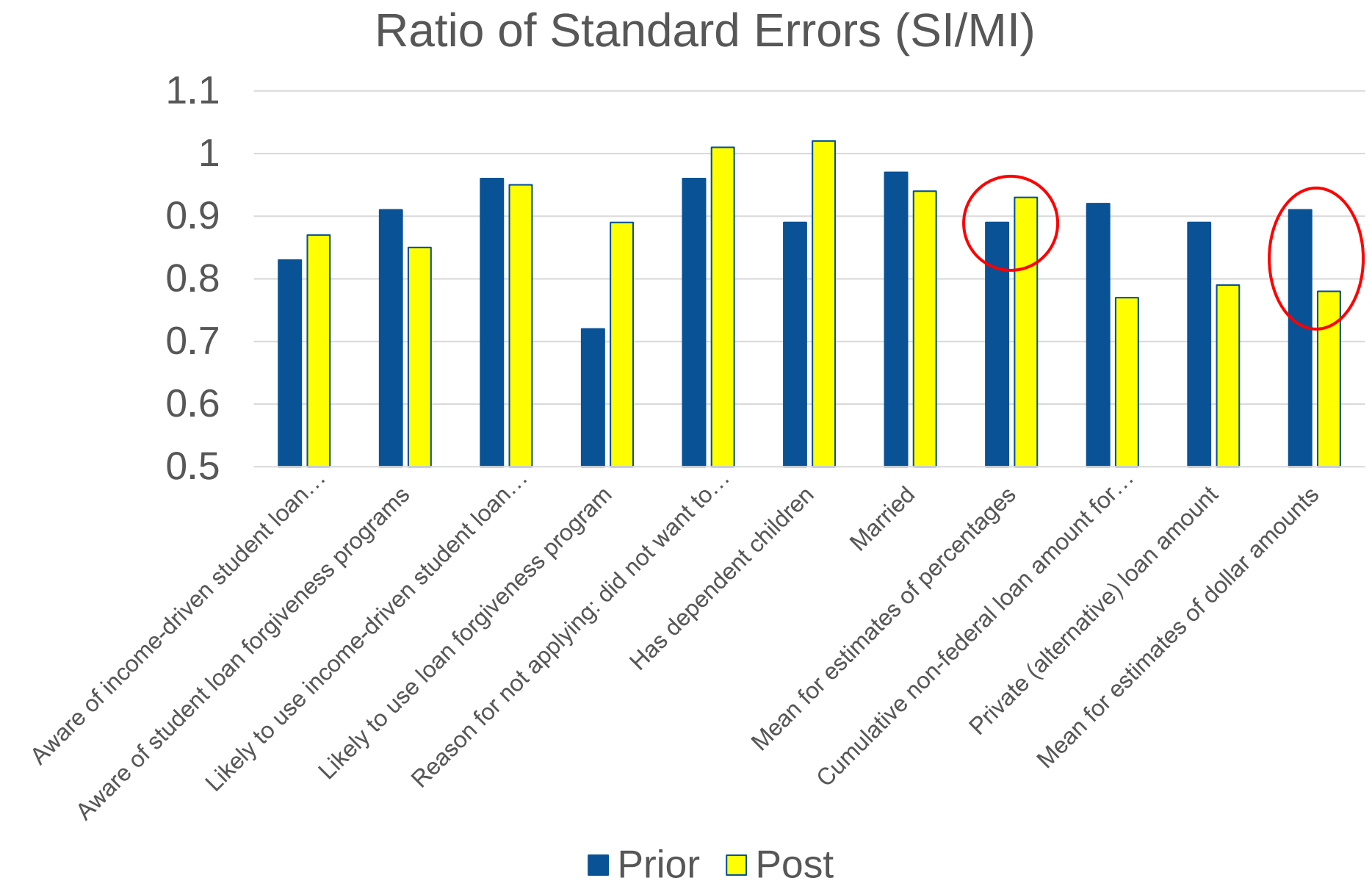
Analysis

- Impact on FMI—maximizing the amount of information
  - Lower is desirable
- Impact on understatement of variance estimates (using single imputation)
  - Ratios of SI/MI closer to one are desirable
- Caveat: in the process of creating a simulated control condition

Results—FMI


0.5
0.4
0.3
0.2
0.1
0
Aware of income-driven student loan repayment…
Aware of student loan forgiveness programs
Likely to use income-driven student loan…
Likely to use loan forgiveness program
Reason for not applying: did not want to take on…
Has dependent children
Married
Mean for estimates of percentages
Cumulative non-federal loan amount for…
Private (alternative) loan amount
Mean for estimates of dollar amounts
FMI Prior to Intervention
FMI Post Intervention

Results—Understatement of Variance Estimates



Summary

- Promising, but:
  - Requires substantial attention to specification of the imputation model(s)
  - Particular attention needed for continuous variables
  - We cannot repeat this test on the same survey because the study member definition has been eliminated

Discussion

- How beneficial would this approach be to a more traditional setting when data are used only from survey respondents and weights to account for nonresponse?
  - Benefit of reducing the understatement of variance estimates is not as pertinent
  - How do we convey the reduction in uncertainty to data users? How do we reflect it in variance estimates?

Thank You

Andy Peytchev
919-541-6648
apeytchev@rti.org