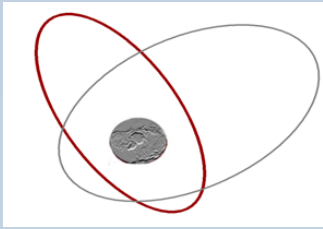
Decomposition of Error Introduced by *Ex-Post* Harmonization



Christopher Ward, University of Michigan
Felicia LeClere, University of Michigan
Pamela Smock, University of Michigan
Lynette Hoelter, University of Michigan
Peter Granda, University of Michigan
James Lepkowski, University of Michigan

ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
A PARTNER IN SOCIAL SCIENCE RESEARCH

Introduction

- Integrated Fertility Survey Series at ICPSR
- Differences over time have consequences for survey quality
- Harmonizing imperfectly comparable variables over time produces error
  - 1. Impact on variable selection
  - 2. Impact on harmonized variable specification
  - 3. Analytical concerns

Situating Harmonization in Total Survey Error

- Proposed model for harmonization error (adaptation of Biemer & Lyberg, 2003):
  - $MSE_H = (B_{SPEC} + B_{NR} + B_{FR} + B_{MEAS} + B_{DP} + B_H)^2 + Var_{SAMP} + Var_{MEAS} + Var_{DP} + Var_H$

where:

- MSE_H = harmonization error-adjusted MSE
- B_H = harmonization bias
- Var_H = harmonization variance

Expansion of Harmonization Error

- Harmonization Bias:
  - $B_H = B_{H\_SPEC} + B_{H\_MEAS} + B_{H\_DP}$
- Harmonization Variance:
  - $Var_H = Var_{H\_SAMP} + Var_{H\_MEAS} + Var_{H\_DP}$
- Harmonization introduces specification, measurement, data processing, and sampling error
  - Impact on quality of data
- How to estimate?
  - Example: specification bias

Specified Harmonized Construct: Number of R's Children in Household

Child Type	1955	1988	1995	2002: NCHILDHH	2002: NUMKDHH
Biological	X	X	X	X	X
Adopted			X	X	X
Step				X	X
Partner's				X	X
Legal ward				X	X
Foster				X	X
Nephew/ niece					X
Grandchild					X

Which Variable to Harmonize?

- NCHILDHH or NUMKDHH?

- Two factors to consider:
  - Minimize the introduction of error
  - Substantive comparability over time

- What we need to know:
  - Which variable overestimates the number of biological or adopted children in the household by a greater margin?
  - Problem: It is difficult to estimate the number of over-counted children

Framework Application

- Five different combinations of types of relationships between the respondent and children in the household

- Let E_1, E_2, E_3, E_4, and E_5 denote:
  - E_1 ={biological}
  - E_2 ={biological, adopted}
  - E_3 ={biological, adopted, stepchild, partner's, legal ward, foster child}
  - E_4 ={biological, adopted, stepchild, partner's, legal ward, foster child, nephew/niece, grandchild}
  - E_5 ={all child relationship types}

- We observe:

Extent of Error

- To estimate the number of miscounted children:

  where:
  E_n is the event of all outcomes (child types) in a given study variable
  E_h is the event of all outcomes (child types) in the harmonized construct

- The total bias depends on the extent to which the children in the sample *do not* belong to *both* events E_n and E_h

- Solution: select NCHILDHH to minimize number of miscounted children in 2002

Specifying the Harmonized Construct

- The number of possible ways to specify the harmonized variable depends on the underlying variables

- Four ways of specifying the harmonized variable (assuming selection of NCHILDHH in 2002):

  - 1) Number of biological children (E_1)
  - 2) Number of biological or adopted children (E_2)
  - 3) Number of biological, adopted, step, partner's, legal ward, or foster children (E_3)
  - 4) Number of all children (E_5)

Consequences for Quality

- The dilemma: how to specify a harmonized variable that both minimizes error and has substantive value to users?
  - Our solution: specify the harmonized variable as "all children"

- The specification bias depends on the extent to which the children in the sample *do not* belong to *both* events E_5 and E_1

Guidelines for Specifying Harmonized Constructs

- Specification bias is unknown but can be estimated
  - Use external data to estimate probabilities

- Specification of harmonized variable depends primarily on two factors:
  - Extent of expected specification bias in a given specification of the harmonized variable
  - Substantive considerations

Summary

- Harmonization error must be considered when harmonizing data *ex-post*
- Example: specification bias influences variable selection, guides specification of the harmonized construct
  - Goal to improve quality of data
- Analytical consequences?
- Generalization of specification bias estimation?

Supplementary Examples

- Highest Grade Attended vs. Highest Grade Completed

- Specifying “Religiously-affiliated” vs. “Church-related” in religious school attendance

- R is Hispanic/Latino
  - Recoded origin vs. direct question