Planned "Missingness" Designs and the American Community Survey (ACS)

Steven G. Heeringa
Institute for Social Research
University of Michigan

Presentation to the National Academies of Sciences Workshop on Respondent Burden in the American Community Survey

March 9, 2016

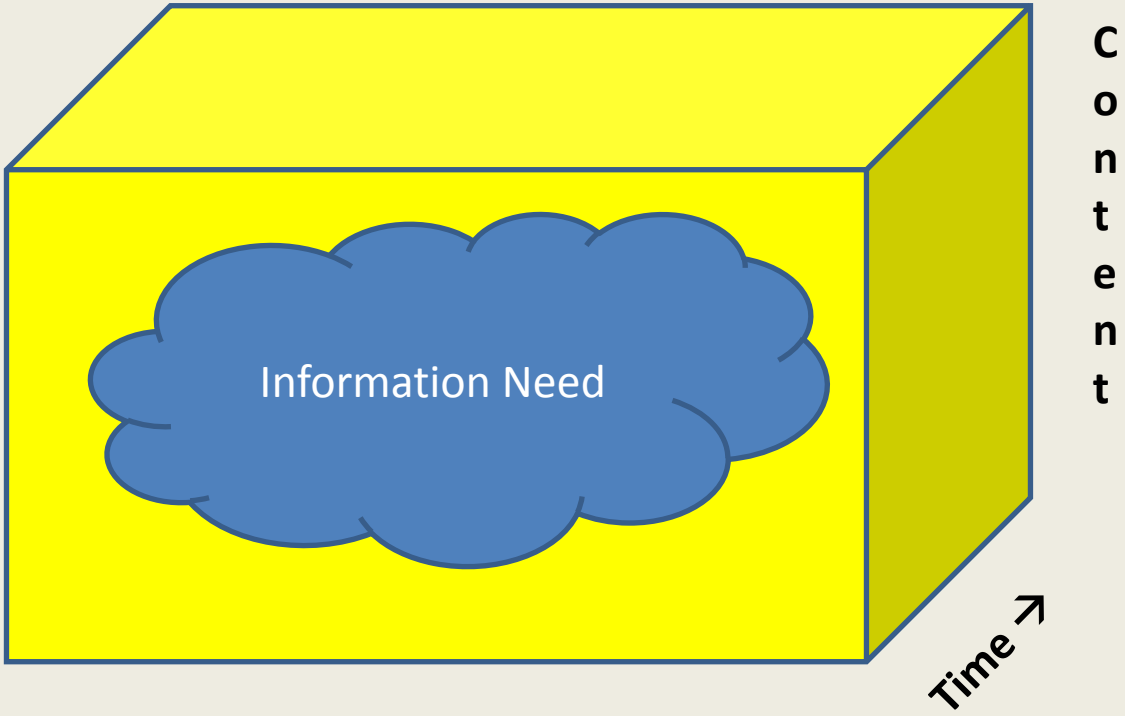
Presentation Outline

- Burden and information needs in the ACS
- Taxonomy of ACS-relevant “planned missingness designs”
- Research-based results, empirical findings and some common sense observations.
- Main ways that planned missingness designs could reduce burden in the ACS.
- Methodological and empirical issues in applying “planned missingness” to the ACS.

Burden Reduction

- Individual Respondent Burden = b(r)
- Aggregate Sample Burden = $\int_r b(\mathrm{r})dr$
- System (Data Producer) Burden
- Data User Burden
- Information Needs as the Driver

3-Dimensions of ACS Information Needs



ACS Information Needs are defined by: 1) Content; 2) Time; and 3) Spatial and Other Domains of Analysis.

ACS Options Considered (Jeff?)

- Option 1: Periodic Inclusion of Questions
- Option 2: Subsampling (multi-phase)
- Option 3: Matrix sampling (SQD)
- Option 4: Administrative records substitution.

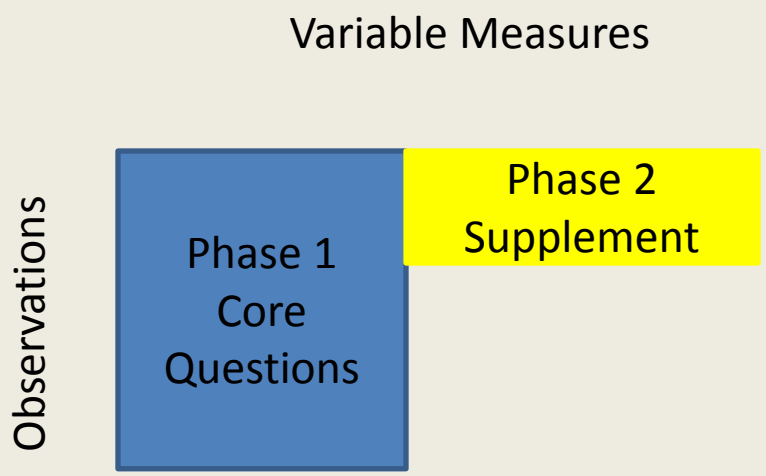
ACS Dimensions for Burden Reduction

- Survey content – Options 3 and 4 and implicitly Option 1.

- Spatial domains – Option 2 and implicitly Option 1
  - ACS is a matrix design over space and time
  - Solution is to collapse design over time and space

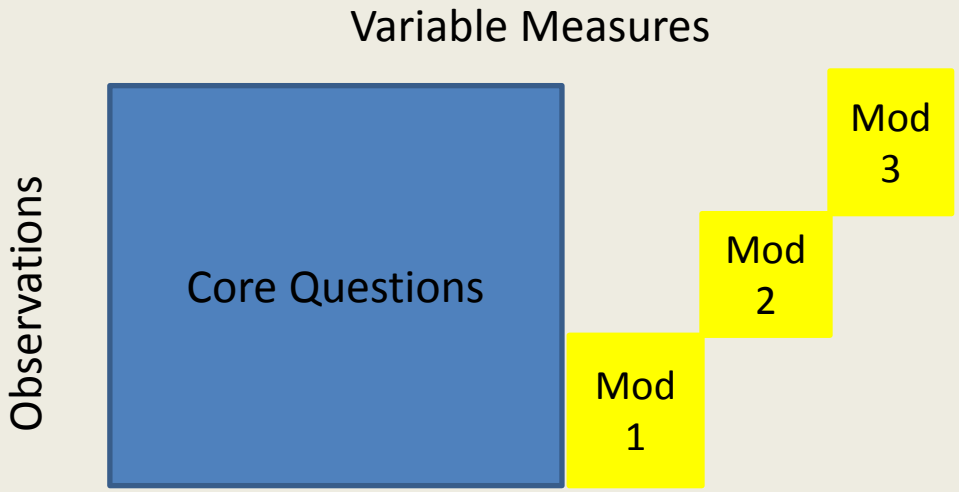
- Time – Option 1

Multi-phase Sampling Design



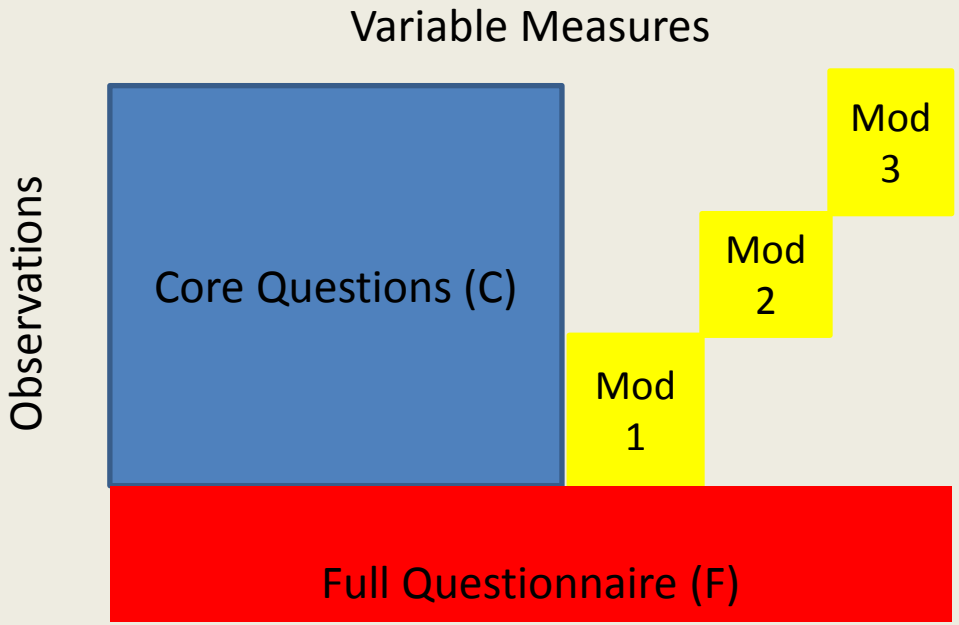
- **Monotonic or nested missing data pattern.**
- **MCAR or MAR mechanism.**
- **Navarro and Griffen (1993)**
  - **Combine with small area estimation methods, e.g. Ericksen (1973)**
  - **Cross tabulations, model-based procedures for generating full sample data. Borrowing strength via empirical Bayes**
- **1970 Census 15% and 5% long form.**
- **National Health and Nutrition Survey (NHANES)**
- **National Comorbidity Survey Replication (NCS-R)**
- **Heeringa (2015)**
- **Little and Rubin (2002)**

Split Questionnaire Design (SQD)



- **Generalized missing data pattern**
- **MCAR or MAR missing data mechanism.**
- **Raghunathan and Grizzle (1995)**
- **Thomas and Gan (1997)**
- **Gonzalez and Eltinge (2007)**
- **Chipperfield, et al. (2013)**
- **Chipperfield and Steele (2009,2011)**
- **Merkouris (2015)**
- **Harel et al. (2015)**

Hybrid Designs



- **Full observation on all variables for part of sample**
- **Full observation on core variables for entire sample**
- **SQD assignment for modular content**
- **Monotonic missing data (F+C) and generalized pattern (F+C+M)**
- **MCAR or MAR missing data mechanism**
- **Chipperfield and Steele (2009)**
- **Merkouris (2015)**

SQD in Practice: What works?

- SQD works in surveys and measurement settings where there are many measures of interest and modularization of correlated items or "blocks" of items is feasible.

- SQD works when the full survey process is designed/redesigned from the ground up for SQD.

- SQD works when estimation requirements are descriptive and predominantly univariate.

Case Study 1: 1985 National School Fitness Survey

- Aim 1: Establish distribution of performance scores (percentiles) on new tests for U.S. children of varying ages.
- Aim 2: Provide data to evaluate comparability of old and new tests in the PYFA program (9 tests total).
- Minimize burden on schools for class time release and students for physical testing across multiple PE classes.

1985 National School Fitness Survey

- Measurement modules: Core survey items and
  - A- (Mile Run, Long Jump, Flex Arm Hang)
  - B- (Pull Ups, 50 yd Dash, Shuttle Run)
  - C- (Two-Mile Walk, Sit and Reach, Sit-ups)
- Module pairs randomized to national probability sample of school classes (i.e. A/B, A/C, C/B)
- Test distributions on 2/3rd of sample
- Test correlations, calibration on 1/3 of sample
- 1985 survey remains standard for participants

Estimation and Inference in Planned Missingness Designs

- Full measurement (F) on a subsample can strengthen the accuracy, precision of estimates and coverage properties of intervals.
- Inclusion of core content (C) for all sample that is correlated with modularized content improves estimation and inference.
- Regardless of strategy (MI, BLUE, CGR), the fraction of missing information is smallest when modularized content has high between module correlation and low within module correlation.

Main ways that planned missingness designs could reduce burden in the ACS.

Option 1: Periodic inclusion of questions

- ACS program has experience with questionnaire changes at the start of the year.
- Annual estimates based on 1 year of data for nation, states and places of N=65,000+.
- For periodic annual estimates, standard weighting, estimation would apply
- Case Study 2 (Time Permitting).

- ACS is a matrix design over space with interpenetrating samples in all but smallest geographies
  - Now: Collapsing over time (i.e. 5 years) support spatial estimates
  - Possible: 5 year rotation of annual topical modules
    - **Collapsing over time (i.e. 5 years) yields a 5 module SQD**

Main ways that planned missingness designs could reduce burden in the ACS.

Option 2: Nested Subsampling for Long Form Administration

- Results in multiple questionnaire forms each year. Complexity in multi-mode administration. Design for stability in long form content across years.

- If systems and protocols for handling multiple forms in the multi-mode ACS data collection could be developed, the subsampling option permits tailoring of subsampling rates to precision requirements for units in the Census geographic hierarchy.

- Greater complexity in processes of editing, imputation and estimation; however, methods for the monotonic data pattern are established and more easily implemented than for the SQD option.

Main ways that planned missingness designs could reduce burden in the ACS.

Option 3: Annual split questionnaire design for content modules

- Likely to be incompatible with complexity and scope of ACS system as noted in *Feasibility Assessment.*

- Amount and type of ACS content that requires annual year-in, year-out measurement does not appear to lend itself well to an annualized SQD measurement, estimation and inference that would truly reduce burden and cost or result in significant information recovery through statistical methods.

Main ways that planned missingness designs could reduce burden in the ACS.

Option 4: Substitution of administrative data

- Obvious direction for continued R & D
- Properties of the data, timeliness, barriers to access and usability may differentially impact estimates over the space, time and content dimensions
- Administrative data sources will not eliminate but may reduce direct survey measurement burden
- Administrative data may play a more important role in model based estimation for small geographic units and short time periods.

Methodological and empirical issues in applying "planned missingness" to the ACS (1).

- Sampling, Questionnaire Design, Data Collection
  - Periodicity and sampling rates for modular/supplemental content (work started for *Feasibility Assessment*)
  - Optimizing core vs. modular question content
  - Control systems for multiple forms and modalities

- Data Processing and Management
  - Adapt processing, editing, imputation and weighting to accommodate the multiple forms and modalities.
  - To what extent do current systems for handling unit nonresponse and item missing data address "missingness by design"

Methodological and empirical issues in applying "planned missingness" to the ACS (2).

- Descriptive statistics for "small" geographical units
  - Composite estimation and small area estimation methods that "borrow strength"
  - Strength can be borrowed from units at the same level or across levels of the Census geography hierarchy
- Crosstabulations, custom tabulation systems for geographical units
  - Available case, EM, FIML, MI?
  - How to use all the statistical information in the incomplete data?
    - Margins, cells, uncertainty bounds
- Public Use Microdata Samples (PUMS), analytic data sets
  - Multiple imputation, fractional imputation
  - User-managed approach to the planned missing data?

Thank you.

sheering@umich.edu

Back Up Slides

Case Study 2: Simple Test of Using ACS Tract-Level Data to Model Small Area Counts of Post 1990 Immigrants to the U.S.

Outcome variable: Post 1990 Foreign Born: the number of population born outside of US 50 states and entered US after 1990.

Target geographical unit: National probability sample area segments comprised of Census Blocks (parts of BGs).

Please note, counts of interest include not only foreign born but also those born in Puerto Rico, U.S. Island Areas or born abroad of American parent(s)). (The latter is 'native born' in ACS terminology)

Synthetic Regression Estimation: Model

Poisson Regression model at the tract level and apply the coefficient to block group covariates to get prediction at the block group.

$$\log(n\_fb1990) = \log(n\_fb) + x' \beta$$

$$= \log(n\_fb) + \beta 0 + \beta 1 x1 + \beta 2 x2 + \ldots\ldots$$

All U.S. tracts excluding 637 with missing data on one or more covariates (e.g. per capita income).

Model estimation cross-validated for 50% sample hold out.

Synthetic Regression Estimation: Predictors*

Census_division

age_18_29 age_30_44 age_45_64 age_ge65

race2_black race2_asian

hh_nonFamily

hhlan_hisp hhlan_asian hhlan_other

income_percap

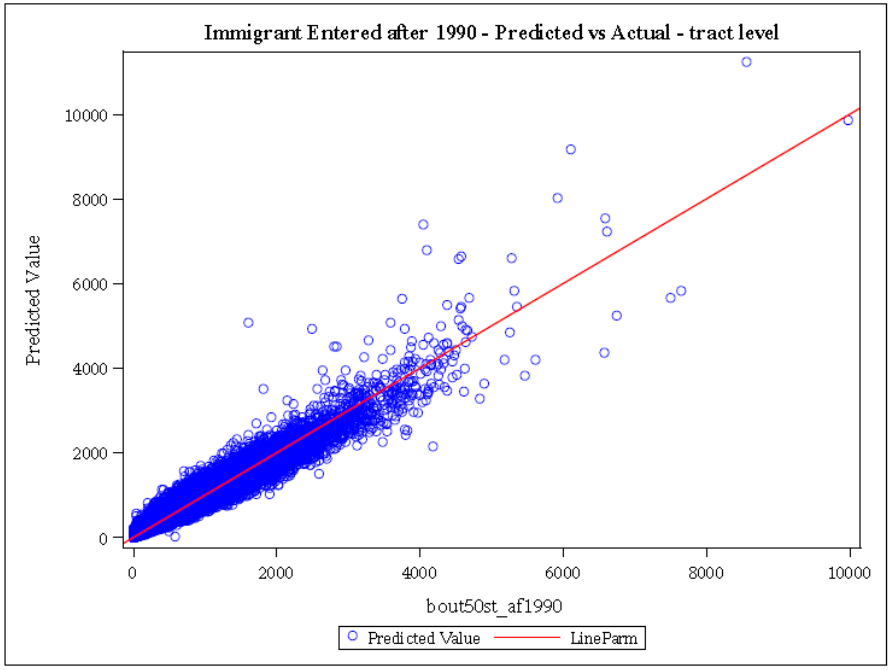
emp_unemplyed emp_other

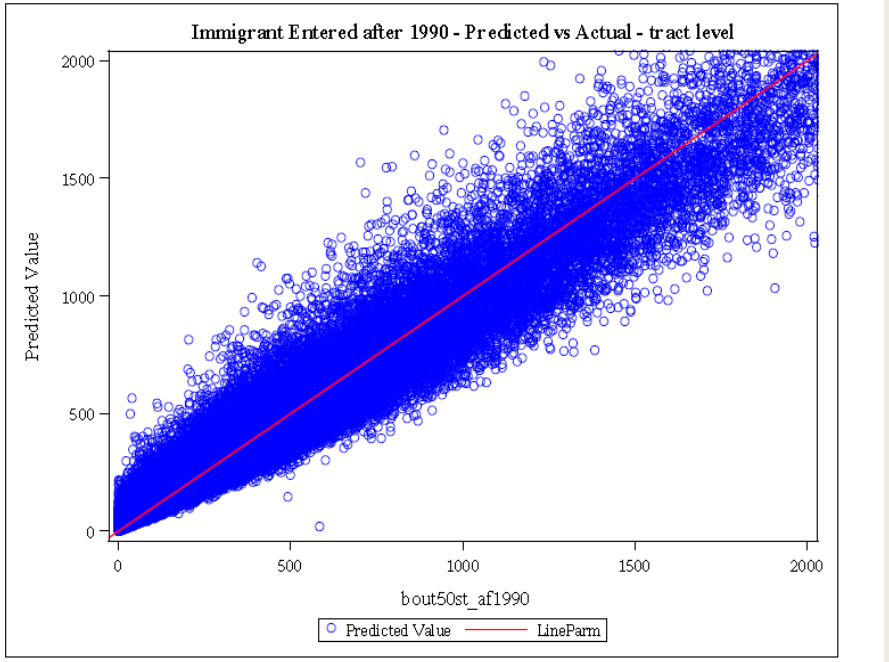
tenure_rent

Foreign_born (offset)

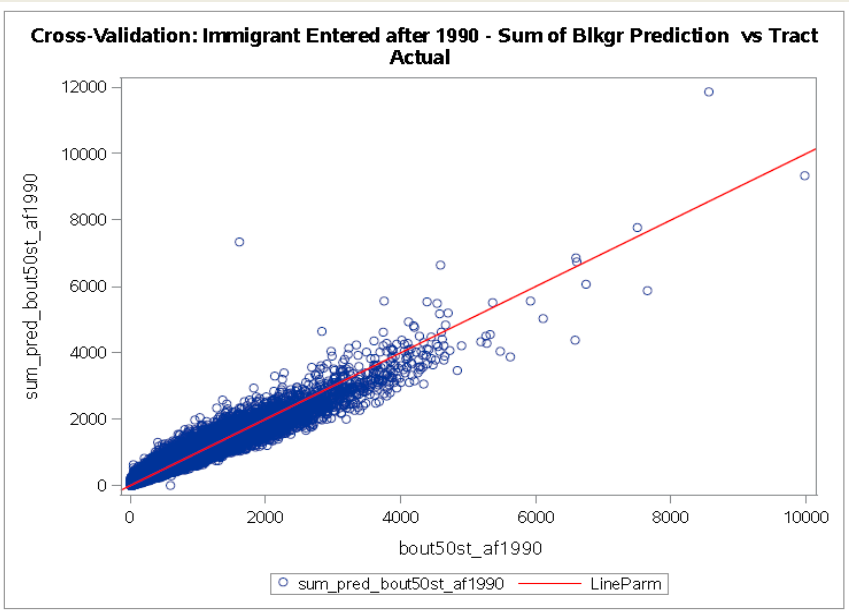
*Available for ACS 5 Year Data at Cenus Tract and Block Group level.

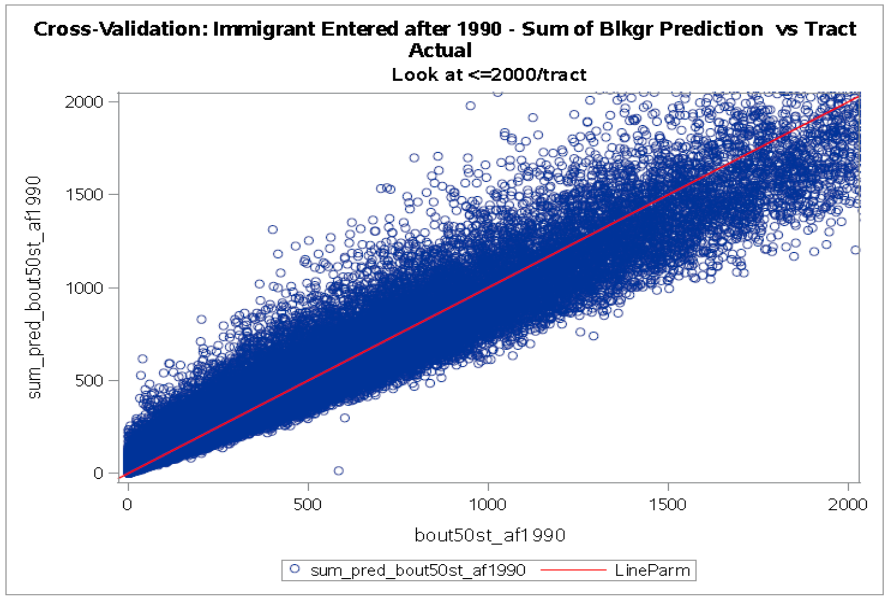
Goodness of fit for Post-1990 Immigrant Pop Model: Offset is Tract Foreign Born Population





Cross-Validation of Syn Reg Model Estimates: Comparing Sum of BG Estimates of Post 1990 Immigrants for a Tract to Actual Tract Values:





Comparing Synthetic Regression HRS 2016 Segment Estimates to ACS 2014 Estimates

Table 1		
ACS2014 5 years	ACS Value	% of population
Population	311,986,080	
bOut50st_af1990	27,917,506	8.95%

Table 2		
Synthetic Regression: HRS2016 Segments Weighted Total (647 Segments)		
	sum	% of population
cen_2010_seg	299,567,264	
seg_est_total_af1990	26,883,964	8.97%
* Table 2 is at person level, calculated by census 2010 population in the segment x estimated prevalence		