



Introduction to The Survey of Income and Program Participation (SIPP)

H. Luke Shaefer
University of Michigan School of Social Work
National Poverty Center

This presentation is part of the NSF-Census Research Network project of the Institute for Social Research at the University of Michigan. It is funded by National Science Foundation Grant No. SES 1131500.

Goals for the SIPP Workshop

- Provide you with an introduction to the SIPP and get you up and running on the public-use SIPP files
- Offer some advanced tools for 2008 Panel SIPP data analysis
- Get you some experience analyzing SIPP data
- Introduce you to the SIPP EHC (SIPP Redesign)
- Introduce you to the SIPP Synthetic Beta (SSB)

H. Luke Shaefer

- Associate professor of social work and public policy (effective September 1, 2015), research affiliate at the National Poverty Center, and one of many co-investigator of the University of Michigan-ISR Census Node
- Interested in how public programs serve low-income families with children
- Began working with the SIPP for my dissertation
 - Still trying to recoup the fixed costs...
- Have worked with public-use files, restricted data, and synthetic data
- Roughly half my published papers use the SIPP
- Proud father of 5-year-old Bridget and 1-year-old Michael

Tell Us About Yourself

- What's your field?
- Where are you at now?
- What attracted you to the workshop?
- What kind of research are you hoping to conduct with the SIPP?

The SIPP

- Nationally representative, longitudinal, multi-stage stratified sample
- Continuous data in 2.5 to 4-year panels from the 1980s through present
- Sample: Civilian, non-institutionalized U.S. households
- **The SIPP's mission:** "to provide a nationally representative sample for: evaluating annual and sub-annual dynamics of income, movements into and out of government transfer programs, the family and social context of individuals and households, and interactions between these items."

The SIPP

- Originally designed to compensate for the limitations of the Current Population Survey (CPS)
 - CPS ASEC (March Supplement) uses a very long recall period
 - Not good at measuring irregular/ odd sources of income
 - High levels of under-reporting of program participation
 - Doesn't capture changes in family structure over time
 - Note: If this makes you panic about the accuracy of our official poverty/insurance estimates from CPS, no-one will blame you
- SIPP was designed to provide month-level detail and have a (much) shorter recall period
- SIPP is meant to provide better estimates of income and public program participation
- Offers the most detailed income and comprehensive program participation variables of the major nationally representative household surveys

What's included in the SIPP?

- The core monthly files include:
 - **Demographics:** race & ethnicity, age, sex, household/ family structure & relationships, state identifiers, education, marital status, student status
 - **Income:** Person/family/household earned income, total income, property income, "other income" unit-specific poverty thresholds, receipt of severance pay
 - **Employment:** Employed, unemployment, not in the labor force, Data on up to 2 jobs/month, industry, occupation, class of worker (public/private), firm size, union membership, tenure, hourly/salaried, employer-based health insurance

What's included in the SIPP?

- The core monthly files include:
 - **Program participation:** Unemployment Insurance, AFDC/ TANF, social security, SSI, SSD, workers' comp, child support, food stamps (SNAP), public housing assistance, energy assistance, public (and private) health insurance, free/reduced lunch, veterans payments, pell grants receipt, other federal grant program receipt
 - Most program participation variables include both receipt (0,1) and the amount of benefit
- **THIS IS NOT A COMPREHENSIVE LIST!**
- <http://www.census.gov/programs-surveys/sipp/tech-documentation/data-dictionaries/data-dictionaries-2008.html>

What's included in the SIPP?

- **Topical Modules:** Extra questions added to the core once per year/panel in particular waves. These include point-in-time/annualized variables on things such as:
 - Fertility history
 - Migration history
 - Material hardship measures (such as food security)
 - Assets and liabilities
 - Medical expenses/utilization of health care
 - Work schedule
 - Note: A major change with the 2014 SIPP redesign will be to incorporate TM content into the core interview
 - Most content areas have been retained, though some at reduced levels of detail
 - Some areas will have expanded detail
- **THIS IS NOT A COMPREHENSIVE LIST!**
- <http://www.census.gov/programs-surveys/sipp/tech-documentation/topical-modules.html>

SIPP Panels: Dates and Sample Size

| Panel | Dates | Wave 1, ref 4 Household Heads | Wave 1, ref 4 n |
|--|-----------|-------------------------------|-----------------|
| 1976-1979 Income Survey Development Program panel: Data are not readily available, but you may be able to get them | | | |
| 1984-1989 panels: harder to access, different file structure—still, they are available and valuable | | | |
| 1990 | 1989-1992 | 21,800 | 58,100 |
| 1991 | 1990-1993 | 14,200 | 37,400 |
| 1992 | 1991-1995 | 19,500 | 51,200 |
| 1993 | 1992-1995 | 19,796 | 52,000 |
| 1996 | 1996-2000 | 36,730 | 95,300 |
| 2001 | 2001-2003 | 35,100 | 90,200 |
| 2004 | 2004-2007 | 43,500 | 110,700 |
| 2008 | 2008-2013 | 42,000 | 105,600 |

Major changes start with the 1996 panel

Sample Selection & Data Editing

- Nationally representative of the US non-institutional population
 - Institutionalized individuals (e.g. those in prison or nursing homes) are not included
 - Also designed to provide reliable estimates at the state level
- Households from areas with high poverty concentrations are oversampled
- Primary wave 1 data collection unit is the **household**, not the **individual**
- Multi-stage stratified sample design:
 - 1) Selection of primary sampling units (PSUs) from strata of similar areas
 - 2) Selection of address units within PSUs

Sample Selection & Data Editing

- The sampling frame is a list of US counties and independent cities, based on the most recent Decennial Census
- Addresses in frames are clustered into “Primary Sampling Units” or PSUs (usually counties or collections of counties)
- PSUs are organized into strata that are similar based on a series of characteristics
- PSUs are then selected from each strata
- Then, addresses are selected from within the selected PSUs

Sample Selection & Data Editing

Data Editing

- When one person in a household is missing for an interview and a “proxy” interview cannot be collected
- Or when particular items are missing for various reasons:
- Public use data are imputed
- Imputation is a sequential hot-deck procedure: missing data are matched with a donor who has similar characteristics (The Redesignated SIPP will have new imputation procedures)
- Income variables are top-coded to protect respondents and edited for logical consistency

Using the Full Panel (Attrition)

- All adult original sample persons (OSP) are followed for the duration of the panel, unless they leave the sample universe
- Attrition is a problem, and is non-random. This matters for those of us that study vulnerable populations
- If an adult (15+) OSP moves to another address, they are followed there, and everyone they live with there is followed—as long as they live with the original sample person
- When an entire household is missing at a wave interview, they are dropped for the wave (but could reappear later)
- **Full panel sample:** Original sample followed for the duration—**must use panel weights for such analyses**

File Structure: Public Use Files

- These bullets pertain to the 2008 and prior panels, and will not be accurate for the new SIPP EHC discussed later in this presentation
- Each panel consists of an independent sample that is followed for 2.5 to 4.5 (ish) years
- Core data are collected every 4 months during waves, reporting on the previous 4 months
 - Not surprising that reporting months are most accurate
- There are 4 randomly selected rotation groups in each wave
- Census staggers the start of each rotation group by month

File Structure (<=2008 Panels)

| Reference Month | Rot Grp 1 | Rot Grp 2 | Rot Grp 3 | Rot Grp 4 |
|-----------------|-----------|-----------|-----------|-----------|
| 12/95 | W1 Ref1 | | | |
| 1/96 | W1 Ref2 | W1 Ref1 | | |
| 2/96 | W1 Ref3 | W1 Ref2 | W1 Ref1 | |
| 3/96 | W1 Ref4 | W1 Ref3 | W1 Ref2 | W1 Ref1 |
| 4/96 | W2 Ref1 | W1 Ref4 | W1 Ref3 | W1 Ref2 |
| 5/96 | W2 Ref2 | W2 Ref1 | W1 Ref4 | W1 Ref3 |
| 6/96 | W2 Ref3 | W2 Ref2 | W2 Ref1 | W1 Ref4 |
| 7/96 | W2 Ref4 | W2 Ref3 | W2 Ref2 | W2 Ref1 |
| 8/96 | W3 Ref1 | W2 Ref4 | W2 Ref3 | W2 Ref2 |
| 9/96 | W3 Ref2 | W3 Ref1 | W2 Ref4 | W2 Ref3 |
| 10/96 | W3 Ref3 | W3 Ref2 | W3 Ref1 | W2 Ref4 |

File Structure

- Core file wave data are organized into person-month observations
 - Each person will have a record for each month they appear in the wave—so up to 4 observations, per person, per wave
- Household/family/subfamily variables are duplicated in each person’s monthly record
- So each household member record contains the household/family/subfamily variables, leading to lots of duplication
 - This can cause confusion (at least for me...)
 - **Example:** RFOKLT18--“Number of own children under 18 in family”

Some Definitions

- **Households:** “a group of persons who occupy a housing unit”
 - Includes: Families, a group of friends sharing a house, two unrelated families, co-housed, an unmarried mother and boyfriend
 - Does not include group quarters: boarding houses, college dorms, monasteries
- **Family:** 2+ people related by birth, marriage, or adoption who reside together
 - See any potential problems here, given family complexity?
- **Related subfamily:** A nuclear family related to but not including the household reference person
- **Unrelated subfamily:** A nuclear family that is not related to the household ref person

Topical Modules

- Come in separate files
- Can merge topical module content into the core using person/household identifier, wave, and reference month variables
- 1996-2008: Topical modules usually attach to the final (4th) reference month of the current wave
- There is a lot of variation with the different topical modules in terms of reference period and known quality of the data
- Comprehensive list available here:
- <http://www.census.gov/programs-surveys/sipp/tech-documentation/topical-modules.html>

The SIPP Synthetic Beta and Restricted data

- In the Census RDC (undisc. location on a number of campuses across the country), with *special sworn status*, it is possible link SIPP with SSA & IRS, and other admin data
 - Can get SIPP data connected to things like lifetime earnings, SSA benefit reciprocity
- It is a priority for Census to make these data available to the public—**while also protecting the privacy of respondents**
- Thus they have created the SIPP Synthetic Beta (SSB)
 - <http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html>
 - Synthesized data based on real micro-data that “should reproduce the characteristics of the underlying confidential micro-data” outside of the RDC
- For the time being, Census will duplicate SSB results on confidential data and share **approved** output

The SIPP Redesign

- Because of concerns about respondent burden, cost, and the complexity of the data, the Census Bureau has just recently redesigned the SIPP
- The NEW SIPP uses an annual recall, but with an event history calendar (EHC)
- First wave is in the field now, reporting on calendar year 2013
- A National Research Council Panel will compare the 2008 panel estimates from 2013 to the new redesigned SIPP
- EHC pilot data are proving to compare well in some regards to the 2008 estimates
- There is overlap between the 2008 SIPP panel and the new SIPP EHC in 2013, which will allow for comparisons
- Still lots of time to use the 2008 panel, which was fielded through much of 2013!

What is the SIPP Good For?

- Can use SIPP as a stacked sample of repeated cross sections
- Can generate monthly national/state-level estimates
- Can generate annualized estimates
- Results appear most accurate for reporting months (reference month 4) in each wave
 - This issue is commonly referred to as “seam bias”
- Estimates must be adjusted for sample design
- The SIPP’s most powerful use is for longitudinal analysis (that’s what will warm my heart to see you do)

What Isn't It Good For?

- Studying the top of the income distribution
- Long longitudinal analyses (over a life course, say, use PSID)
- If you need annual estimates for every year
- If you don't want to deal with the complexities in household/family composition that the SIPP uncovers...

Use the SIPP When

- You want to deal with more of the complexity of messy questions
- You want the best available survey-based estimates of the income of the poor
- You want to benefit from overall higher reporting rates for public program participation
- You want to conduct longitudinal analyses over relatively short periods (month-to-month; annualized; up to 4 years)

The Uninsured in America, 2005 Non-Elderly (Estimates by ERIU)

| Dataset | During the year | All year | Point-in-time |
|---------|-----------------------|-----------------------|-----------------------|
| CPS | N/A | 44.4 million 17.2% | N/A |
| SIPP | 65.9 million 25.6% | 27.6 million 10.7% | 45.2 million 17.6% |
| MEPS | 66.9 million 25.7% | 36.5 million 14.0% | 50.1 million 19.2% |

Health Insurance & Divorce

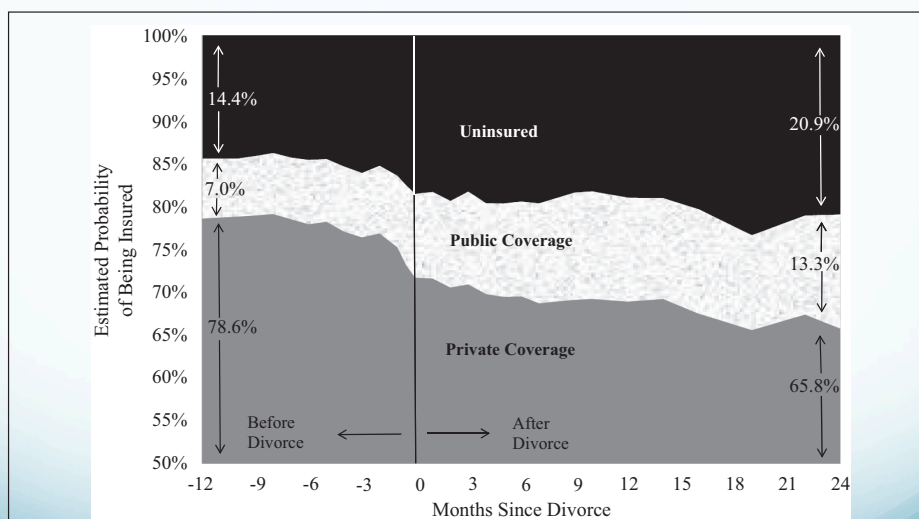


Figure 1. The Changing Distribution of Women's Health Insurance Coverage across Time since Divorce

Lavelle, B. & Smock, P.J. (2012). Divorce and Women's Risk of Health Insurance Loss. *Journal of Health and Social Behavior*, 53(4), 413-431.

Workshop Resources

- Data files for 2008 available core public use waves (in stata)
- Data files for some key topical modules
- Version of the current user guide with updated chapters merged in (page numbers will be off)
 - <http://www.census.gov/programs-surveys/sipp/methodology/users-guide.html>
- Some useful technical papers on the SIPP
- Exercises that you might find handy to get to know the data