Survey of Income and Program Participation

Jason Fields
Matthew Marlay
Patrick Campanello

US Census Bureau

December, 2015 – FCSM
Washington, DC
Outline

- Reengineering SIPP background and key changes
- Timeline and release plans
- Staffing and training
- Some innovations
  - Model based incentives
  - Model based imputation with administrative inputs
  - Monitoring
- New data access tool from Orlin Research
Survey of Income and Program Participation

- National panel survey – Since 1984 with sample size between about 11,000 and 45,000 interviewed households
  - The duration of each panel varies from 2½ yrs to 4 yrs
  - The SIPP sample is a multistage-stratified sample of the U.S. civilian non-institutionalized population
  - Longitudinal – following original sample household members (all 15 and over are followed)

- SIPP ‘Classic’ – 1984-2008
  - Uses a 4-month recall period – 3 interviews / year
  - The sample is divided into 4 rotation groups for monthly interviewing
  - Paper from 1984-1993 and DOS based CASES instrument from 1996-2008

- SIPP 2014
  - Annual survey with four month interviewing window – recall to beginning of prior (reference) year
  - Event History Calendar (EHC) component to facilitate recall
  - Paradata and ancillary data include (contact history and reluctance, training certification, keystroke files, cost and case management, prior wave data for waves 2+)

- Interviews are conducted by personal visit and by decentralized telephone if requested for follow-up
Key Design Changes and Benefits of Reengineering

- Annual interview
- 12-month reference period from 4-month
- Event History Calendar (EHC) methods - Facilitates respondent recall over longer reference period
- Reduced cost through annual administration

Scope

- Similar to SIPP
- Broader than core / includes key topical module content in each wave

Better integration of concepts

- EHC - integrates reporting across domains – incorporates dependent data
- Topics previously implemented as add-on modules now integrated

Increased efficiency in processing and producing data products
Flexibility in administration (dynamic interview month and reference period)
Release Plans

- Wave 1 (Collected Feb-Jun 2014)
  - Research file release – *approximately* the end of CY 2015 (limited content)
  - Full public use release – mid-2016
  - Available for RDC use later in 2016

- 2014 Social Security Administration Supplement on Retirement, Pensions and Related Content
  - Dependent on Wave 1 edited input
  - Reviewing WebCATI outcomes to retain additional cases

- Wave 2 (Collected Feb-May 2015)
  - Full public use release – late 2016

- Wave 3 (Fieldwork begins Feb April 2016)
2014 SIPP: Content Overview

- Coverage Questions
- Roster
  - Sex
  - Birthdate/Age
- Demographics
  - Hispanic origin
  - Race
  - Citizenship
  - Language
  - Marital status
  - Parent/child relationships
  - Educational attainment
  - Armed forces status
  - Type 2 people
  - Program/income screeners
- Event History Calendar
  - Residency
  - Marital history
  - Educational enrollment
  - Jobs/Time not working
  - Program receipt
  - Health insurance

- Post-EHC Questions
  - Health insurance
  - Dependent care
  - *Non-job income
  - *Program income
  - Asset ownership
  - Household expenses
  - Health care utilization
  - Medical expenditures
  - Disability
  - *Fertility history
  - *Biological parents’ nativity and mortality
  - Child care
  - Child well-being
  - Adult well-being

- Closing Screens (not on public-use file)
  - Respondent Identification Policy
  - Contact information
  - Moving intentions

Bold – in Research File Release
* – Tentatively included
SIPP 2014 Interviewer Training

- Decentralized training after centralized ‘Train-the-Trainer’ at Census HQ

- Two-day generic Census training
  - New hires only
  - Communicating with respondents
    - Covers cross-survey skills
    - Administrative training

- Four-day classroom training
  - All SIPP Interviewers (FRs)
  - Decentralized verbatim training
  - Paired-practices
    - Content specific to SIPP
    - Daily quizzes
    - Computer based training sequences

- Pre- and post-classroom self-study modules

- Ends with certification test
  - Required before fieldwork can be started
Innovations

- Focused use of dependent data in an Event History Calendar
- Model based incentive assignment
- Type-Z model-based imputation
  - informed by administrative records
  - operationalizing methods discussed in the early 1990s - sequential regression multiple imputation
- Monitoring
  - Integration of paradata streams for management and evaluation
  - Intensive interviewer training – many aspects to monitor
  - CARI – Audio Recorded Interviews
## Completed Cases by Incentive Receipt Wave 1

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Completed Cases</th>
<th>Type As</th>
<th>Type Bs</th>
<th>Type Cs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>53,070</td>
<td>70.16%</td>
<td>29.84%</td>
<td>16.07%</td>
<td>3.93%</td>
</tr>
<tr>
<td>Group 1 ($0)</td>
<td>13,549</td>
<td>69.01%</td>
<td>30.99%</td>
<td>16.16%</td>
<td>4.15%</td>
</tr>
<tr>
<td>Group 2 ($0)</td>
<td>13,471</td>
<td>69.07%</td>
<td>30.93%</td>
<td>16.14%</td>
<td>3.88%</td>
</tr>
<tr>
<td>Group 3 ($20)</td>
<td>13,470</td>
<td>70.22%</td>
<td>29.78%</td>
<td>16.49%</td>
<td>4.10%</td>
</tr>
<tr>
<td>Group 4 ($40)</td>
<td>12,580</td>
<td>72.49%</td>
<td>27.51%</td>
<td>15.44%</td>
<td>3.51%</td>
</tr>
</tbody>
</table>

### Response Improvement by Incentive Group

- **Overall:** 70.16%
- **Group 1 ($0):** 69.01%
- **Group 2 ($0):** 69.07%
- **Group 3 ($20):** 70.22%
- **Group 4 ($40):** 72.49%
### Completed Cases by Incentive Receipt Wave 2

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Completed Cases</th>
<th>Type As</th>
<th>Type Bs</th>
<th>Type Cs</th>
<th>Type Ds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>35,530</td>
<td>74.19%</td>
<td>21.94%</td>
<td>0.68%</td>
<td>11.87%</td>
<td>3.87%</td>
</tr>
<tr>
<td>W1=$0, W2=$0</td>
<td>8,805</td>
<td>72.68%</td>
<td>23.43%</td>
<td>0.77%</td>
<td>11.48%</td>
<td>3.90%</td>
</tr>
<tr>
<td>W1=$0, W2=$40</td>
<td>8,906</td>
<td>76.38%</td>
<td>19.86%</td>
<td>0.63%</td>
<td>12.09%</td>
<td>3.76%</td>
</tr>
<tr>
<td>W1=$20, W2=$0</td>
<td>8,911</td>
<td>73.05%</td>
<td>23.33%</td>
<td>0.67%</td>
<td>12.01%</td>
<td>3.62%</td>
</tr>
<tr>
<td>W1=$40, W2=$0</td>
<td>4,408</td>
<td>73.63%</td>
<td>21.78%</td>
<td>0.73%</td>
<td>11.80%</td>
<td>4.59%</td>
</tr>
<tr>
<td>W1=$40, W2=$40</td>
<td>4,500</td>
<td>75.66%</td>
<td>20.50%</td>
<td>0.60%</td>
<td>12.02%</td>
<td>3.84%</td>
</tr>
</tbody>
</table>

#### Graph: Response Overall and Improvement by Incentive Group

- **Overall**: 74.19%
- **W1=$0, W2=$0**: 0.00%
- **W1=$0, W2=$40**: 0.37%
- **W1=$20, W2=$0**: 0.95%
- **W1=$40, W2=$0**: 2.98%
- **W1=$40, W2=$40**: 3.70%
Wave 3 Incentive Plans

- Incentive assignment model:

  - Logistic regression model that predicts the probability of response using household characteristics such as:
    - Metropolitan status
    - Age
    - Sex
    - Household size
    - Tenure
    - Poverty strata

  - Assign incentives to households with the lowest likelihood of responding without an incentive and highest increase in response given an incentive
Topic Model Imputation

Problem:
- How to improve process for creating fully imputed data where whole people are missing from the household?
  - Previously relied on matching to donors and substituting prior to edits.
  - How to implement new imputation methods and still release data in a timely manner for a survey with 11,000 collected and 2,000 released variables?

Solution
- Replace item-level hot deck with parametric model-based approach
  - Helps handle small hot deck cell size problems
  - Allows inclusion of many more predictor variable
  - SIPP SSB provides the methodological foundation for modelling
  - Use administrative data to mitigate problems caused when survey data are not “missing at random”
- Use topic flags as alternative to whole-record donation for cases where respondent did not complete the whole sections of the survey.
- Indicator variables for all the major topics covered by SIPP (See Ref. Sect. 1)
- Implement new methods only for these 40+ variables
List of Topic Flags in 2014 SIPP

**EHC topics:**
- Education Enrollment
- Employment (job lines 1-7)
- General Assistance
- SNAP
- SSI
- TANF
- WIC
- Health insurance
  - Private
  - Medicaid
  - Medicare
  - Military
  - Other

**Non-EHC topics:**
- Biological parent (fertility)
- Dependent care
- Disability- adult and child functional limitations (seeing, hearing, etc.)
- Disability (difficulty finding or keeping a job because of disability)
- Disability (not being able to work because of disability)
- Disability payments
- Energy assistance
- Lump sum payments
- Retirement
- Retirement payments
- Life insurance
- School lunch
- School breakfast
- Social Security- Adults
- Social Security- Kids
- Survivor payments
- Unemployment compensation
- Veterans affairs benefits
- Worker's compensation
### Results

**Overall Percentages for cases where SIPP respondent answered the first question about jobs held (94.5% of in-universe respondents)**

<table>
<thead>
<tr>
<th>Worked for pay in 2013?</th>
<th>W-2/Schedule C positive earnings in 2012?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>58.2</td>
</tr>
<tr>
<td>No</td>
<td>41.8</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td><strong>58.1</strong></td>
</tr>
<tr>
<td><strong>No</strong></td>
<td><strong>41.9</strong></td>
</tr>
</tbody>
</table>

**Overall Percentages for cases where SIPP respondent DID NOT answer the first question about jobs held and TF was imputed (5.5% of in-universe respondents)**

<table>
<thead>
<tr>
<th>Worked for pay in 2013?</th>
<th>W-2/Schedule C positive earnings in 2012?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>61.5</td>
</tr>
<tr>
<td>No</td>
<td>38.5</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td><strong>60.4</strong></td>
</tr>
<tr>
<td><strong>No</strong></td>
<td><strong>39.6</strong></td>
</tr>
</tbody>
</table>
Topic Model

Conclusions:

- Model-based imputation is feasible in a production environment for a large-scale survey
- Outside data sources (especially administrative data) are valuable:
  - Additional predictor variables in a model
  - Independent of survey non-response mechanism

Next steps:

- Model respondent-reported earnings
- Model beginning and end of spells
  - Help mitigate seam bias
- Model more topics
  - Defined benefit pension contributions
- How to best take account of spouse/parent/sibling relationships in the data when modeling
Paradata/Auxiliary Sources in Use

- Audit trail data from the Blaise/C# instrument
- Certification test for interviewer training
- Interviewer characteristics
  - Census experience
  - Prior SIPP experience
  - Supervisory status
  - Demographics
- Contact history instrument
- Mileage, case load, supervisor observation, hours billed
- Neighborhood observation
- Regional office progress management application data
- Interviewer debriefing
- Interview recordings
New data access tool from Orlin Research

- The Orlin Longitudinal Data System (OLDS) is a tool that organizes SIPP data, creating the necessary linkages across persons and units and over time.

- It allows the analyst to easily recode variables and manage data using these linkages and includes full linkages with SIPP metadata such as questionnaires and variable descriptions:
  - Data and metadata search and exploration tools
  - Automatic tracking of relationships across records types and across time
  - Easy variable creation and modification

- Built-in analysis tools, using the R statistical language:
  - Simple templates provided for each type of data manipulation or analysis
  - Complete audit trail, documenting all actions performed
  - Export of data in any format

- Orlin Tool has loaded 1996-2008 SIPP Panels:
  - The 2008 SIPP panel: 60 months of information on 131,337 individuals
  - Core data and topical modules
  - Linked longitudinally at the person level
Introduction – The Interface

**Browse**
Search all metadata or explore specific items by selecting them in the metadata tab.

**Explore**
Select a data collection to browse in the data tab. View data linked to codes and frequencies.

**Export**
Export data in many formats for further analysis in statistical packages.

**Transform**
Recode and restructure your data advanced capabilities for linking records.

**Edit**
Collaborative metadata editing allow tracked and coordinated editing of metadata.

**Analyze**
Analyze your data with the embedded R Statistical package or export to a program of your choice.
Continuous Variables
Let's Recode RMESR

Recode: person_months: RMESR ->
person_months.RMESR_RECODE

Existing values for variable RMESR:

<table>
<thead>
<tr>
<th>Count</th>
<th>Value</th>
<th>Label</th>
<th>New Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,022,097</td>
<td>-1</td>
<td>Not in universe</td>
<td>0</td>
</tr>
<tr>
<td>2,222,905</td>
<td>1</td>
<td>With a job entire month, worked all weeks</td>
<td>1</td>
</tr>
<tr>
<td>46,785</td>
<td>2</td>
<td>With a job entire month, absent from work without pay 1+ weeks, absence not due to layoff</td>
<td>1</td>
</tr>
<tr>
<td>31,151</td>
<td>3</td>
<td>With a job entire month, absent from work without pay 1+ weeks, absence due to layoff</td>
<td>1</td>
</tr>
<tr>
<td>23,335</td>
<td>4</td>
<td>With a job at least 1 but not all weeks, no time on layoff and no time looking for work</td>
<td>1</td>
</tr>
<tr>
<td>24,373</td>
<td>5</td>
<td>With a job at least 1</td>
<td>1</td>
</tr>
</tbody>
</table>

New Codes for the variable person_months.RMESR_RECODE

1: working

- Code type: ○ values ○ ranges ○ all uncoded
- Values: 
- Range: 1 5
- This code is: □ a missing value □ an invalid value

0: Not working
And we get the stats for the new var

<table>
<thead>
<tr>
<th>Code</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not working</td>
<td>2767096</td>
</tr>
<tr>
<td>Working</td>
<td>2346549</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Maximum</td>
</tr>
<tr>
<td>Standard Deviation</td>
</tr>
<tr>
<td>Valid Cases</td>
</tr>
<tr>
<td>Missing Cases</td>
</tr>
</tbody>
</table>
With the changes pointed out:

```r
use(SIPP,'2008',subfamily_months);

create_variable(sf_tanf_owner,'Subfamily tanf owner','person_months.RCUOWN20');
create_variable(sf_ref_person,'Subfamily reference person number','person_months.sf_ref_person');

create_variable(sf_num_kids,'Number of children in this subfamily this month','count(person_months,TAGE < 16)');
create_variable(sf_age_head,'Age of the head','max(person_months.TAGE,EPPNUM EQ sf_ref_person)');
create_variable(sf_work_head,'Employment status of head','max(person_months.RMESR_RECODE,EPPNUM EQ sf_ref_person)');
create_variable(sf_on_tanf,'Is this subfamily not on TANF this month','case when sf_tanf_owner > 0 then 1 else 0');
create_variable(sf_on_tanf,'Is this subfamily on TANF this month','case when sf_tanf_owner > 0 then 1 else 0');
```

Close
Create a regression - Regression for sf_on_tanf_u13

Call:
glm(formula = sf_on_tanf_u13 ~ sf_num_kids_u13 + sf_age_head_u13 +
    sf_work_head_u13, family = binomial(link = "probit"), data = y)

Deviance Residuals:
          Min         10       Median       30       Max
-0.9401    -0.2005     -0.1566    -0.1401     3.3617

Coefficients:
                         Estimate      Std. Error  z value  Pr(>|z|)
(Intercept)            -1.3404158     0.0084722  -158.21 <2e-16 ***
sf_num_kids_u13        -0.0668372     0.0010849   -61.60 <2e-16 ***
sf_age_head_u13        -0.0090036     0.0002112   -42.63 <2e-16 ***
sf_work_head_u13       -0.7772969     0.0037101  -209.51 <2e-16 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 567999   on 2253572  degrees of freedom
Residual deviance: 512142   on 2253560  degrees of freedom
(128479 observations deleted due to missingness)
AIC: 512150

Number of Fisher Scoring iterations: 7
And the chart
### Transitions

**2008: person_months**

<table>
<thead>
<tr>
<th>Person longitudinal key</th>
<th>Reference month of this record</th>
<th>Total Family public assistance payments</th>
<th>Total related subfamily public assistance payments</th>
<th>Longitudinal month</th>
<th>Start TANF prev</th>
<th>Is this subfamily not on TANF this month</th>
<th>household_month</th>
</tr>
</thead>
<tbody>
<tr>
<td>20343001</td>
<td>Fourth reference month</td>
<td>362</td>
<td>None or not in universe</td>
<td>16</td>
<td>true</td>
<td>1</td>
<td>household_month</td>
</tr>
<tr>
<td>20343001</td>
<td>First reference month</td>
<td>537</td>
<td>None or not in universe</td>
<td>17</td>
<td>false</td>
<td>0</td>
<td>household_month</td>
</tr>
<tr>
<td>20343001</td>
<td>Second reference month</td>
<td>537</td>
<td>None or not in universe</td>
<td>18</td>
<td>false</td>
<td>0</td>
<td>household_month</td>
</tr>
<tr>
<td>20343001</td>
<td>Third reference month</td>
<td>537</td>
<td>None or not in universe</td>
<td>19</td>
<td>true</td>
<td>1</td>
<td>household_month</td>
</tr>
<tr>
<td>20343001</td>
<td>Fourth reference month</td>
<td>537</td>
<td>None or not in universe</td>
<td>20</td>
<td>false</td>
<td>1</td>
<td>household_month</td>
</tr>
<tr>
<td>20343001</td>
<td>First reference month</td>
<td>382</td>
<td>None or not in universe</td>
<td>21</td>
<td>false</td>
<td>1</td>
<td>household_month</td>
</tr>
<tr>
<td>20343001</td>
<td>Second reference month</td>
<td>362</td>
<td>None or not in universe</td>
<td>22</td>
<td>false</td>
<td>1</td>
<td>household_month</td>
</tr>
<tr>
<td>20343001</td>
<td>Third reference month</td>
<td>362</td>
<td>None or not in universe</td>
<td>23</td>
<td>false</td>
<td>1</td>
<td>household_month</td>
</tr>
</tbody>
</table>
Creating a spell

Create a spell record -

- **Rec type**: person_months
- **Time variable**: person_months.LGTMON
- **Observation unit**: persons
- **Minimum spell duration**: 1
- **Spell gap tolerated**: 0
- **Spell name**: tanf
- **Spell label**: Tanf Spell
- **In spell expression**: sf_on_tanf=1
- **Not in spell expression**: sf_on_tanf=0
- **Case selection**: 
- **Vars to copy**: sf_age_head, sf_fam_type, sf_work_head, sf_family_kind
THANK YOU!

Jason Fields  
Survey Director  
Survey of Income and Program Participation  
National Survey of Children’s Health  
Jason.M.Fields@census.gov

www.census.gov  
www.census.gov/sipp