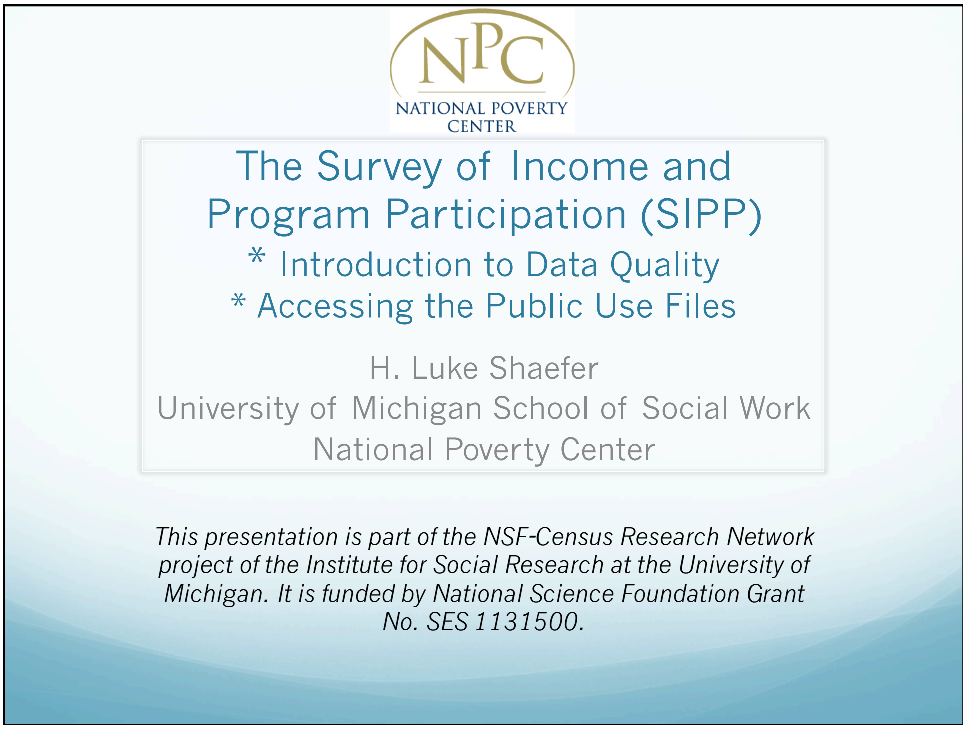


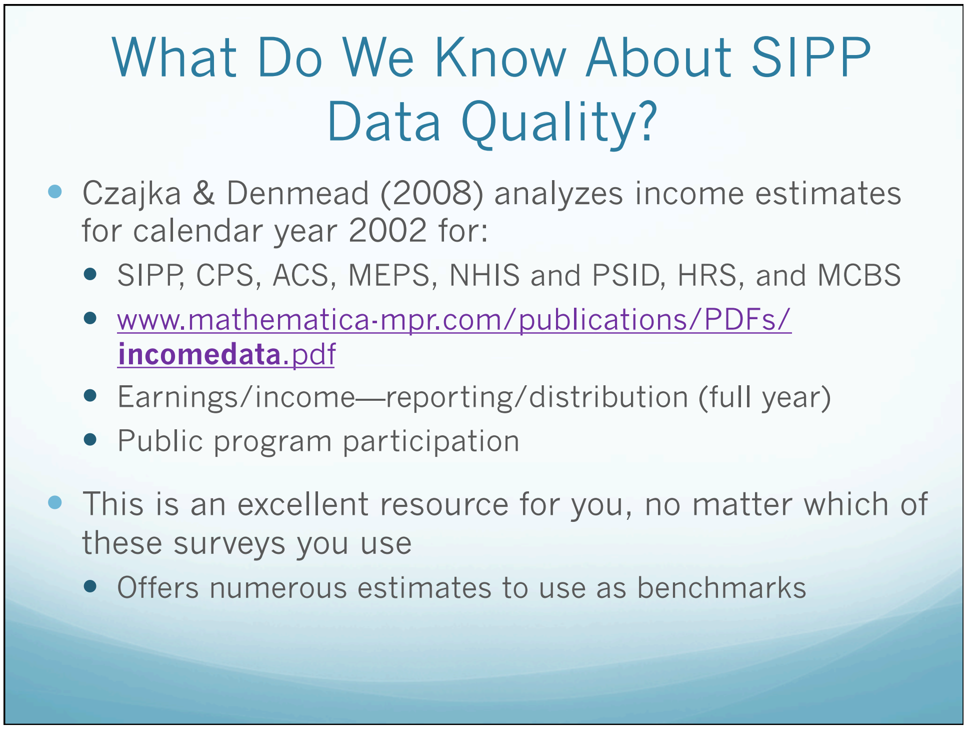
NATIONAL POVERTY CENTER

# The Survey of Income and Program Participation (SIPP)

## * Introduction to Data Quality
## * Accessing the Public Use Files

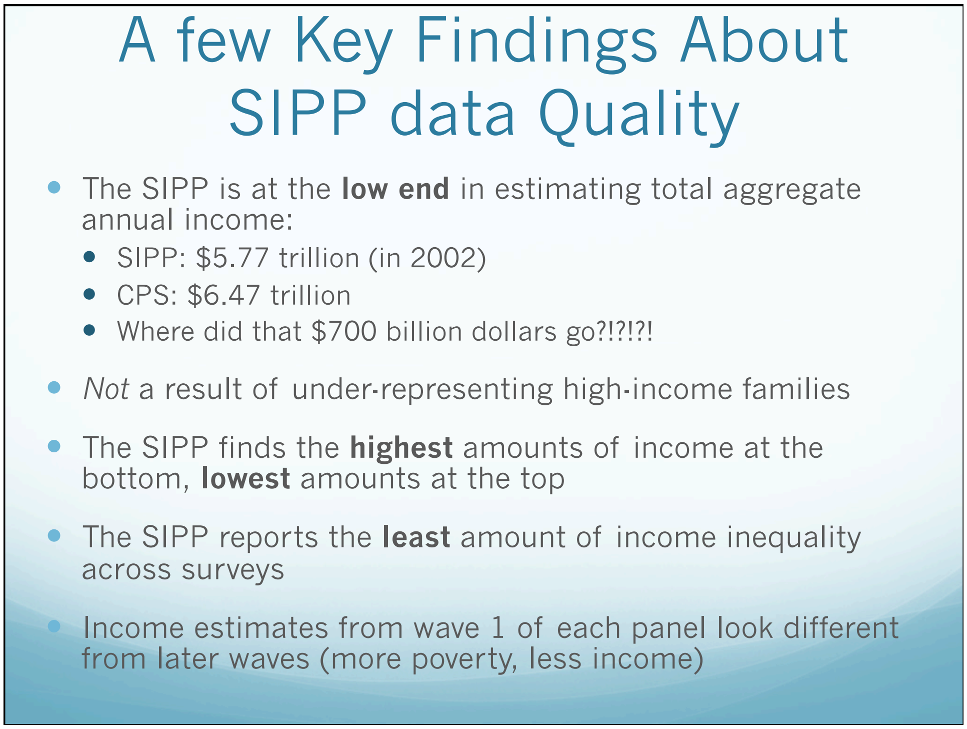
H. Luke Shaefer
University of Michigan School of Social Work
National Poverty Center

*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.*



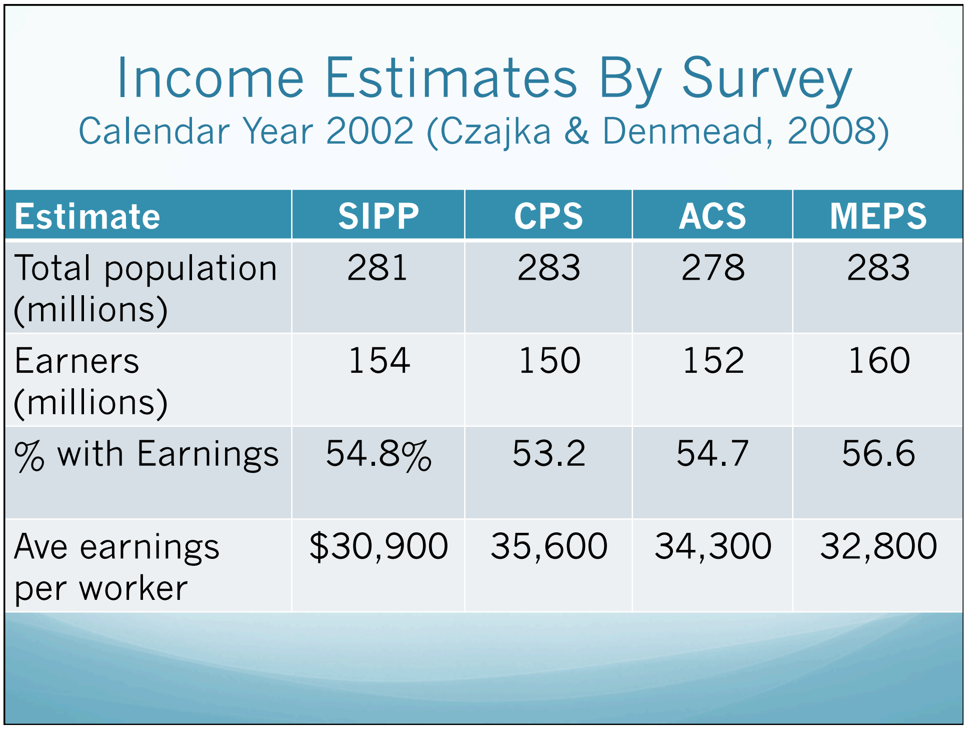
# What Do We Know About SIPP Data Quality?

- Czajka & Denmead (2008) analyzes income estimates for calendar year 2002 for:
  - SIPP, CPS, ACS, MEPS, NHIS and PSID, HRS, and MCBS
  - www.mathematica-mpr.com/publications/PDFs/**incomedata**.pdf
  - Earnings/income—reporting/distribution (full year)
  - Public program participation
- This is an excellent resource for you, no matter which of these surveys you use
  - Offers numerous estimates to use as benchmarks



# A few Key Findings About SIPP data Quality

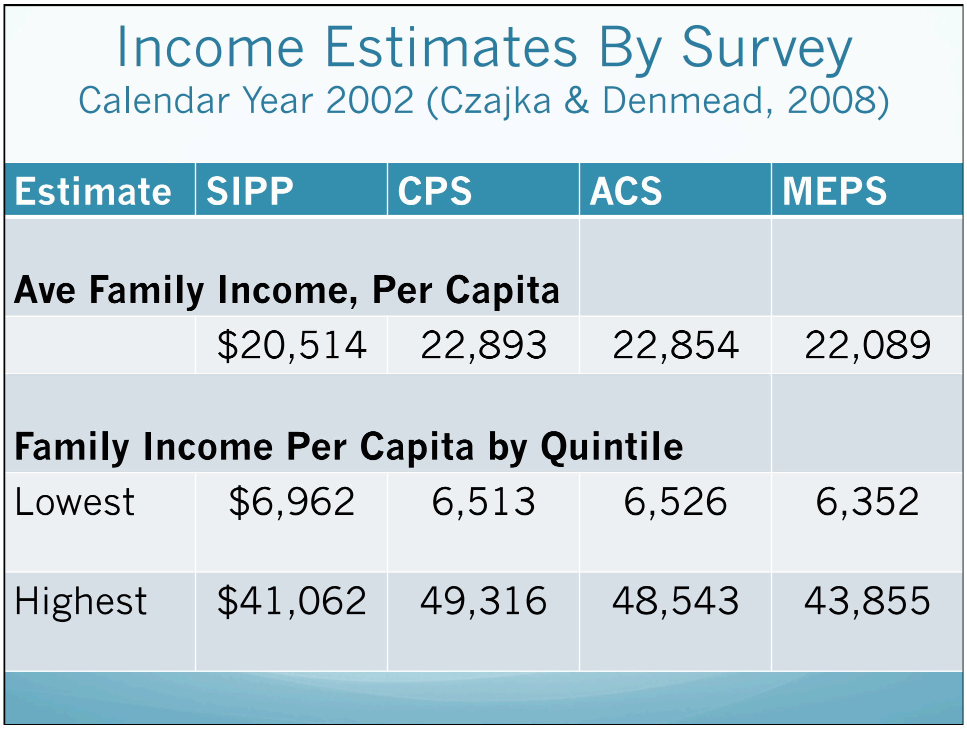
- The SIPP is at the **low end** in estimating total aggregate annual income:
  - SIPP: $5.77 trillion (in 2002)
  - CPS: $6.47 trillion
  - Where did that $700 billion dollars go?!?!?!
- *Not* a result of under-representing high-income families
- The SIPP finds the **highest** amounts of income at the bottom, **lowest** amounts at the top
- The SIPP reports the **least** amount of income inequality across surveys
- Income estimates from wave 1 of each panel look different from later waves (more poverty, less income)



# Income Estimates By Survey

Calendar Year 2002 (Czajka & Denmead, 2008)

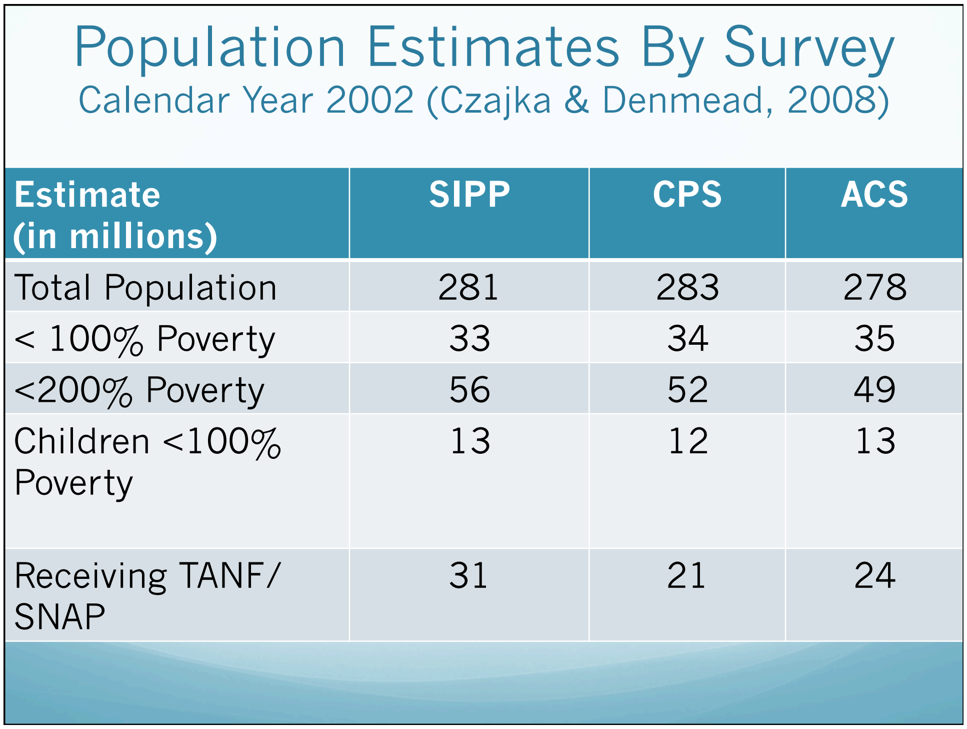
| Estimate | SIPP | CPS | ACS | MEPS |
|---|---|---|---|---|
| Total population (millions) | 281 | 283 | 278 | 283 |
| Earners (millions) | 154 | 150 | 152 | 160 |
| % with Earnings | 54.8% | 53.2 | 54.7 | 56.6 |
| Ave earnings per worker | $30,900 | 35,600 | 34,300 | 32,800 |



## Income Estimates By Survey

Calendar Year 2002 (Czajka & Denmead, 2008)

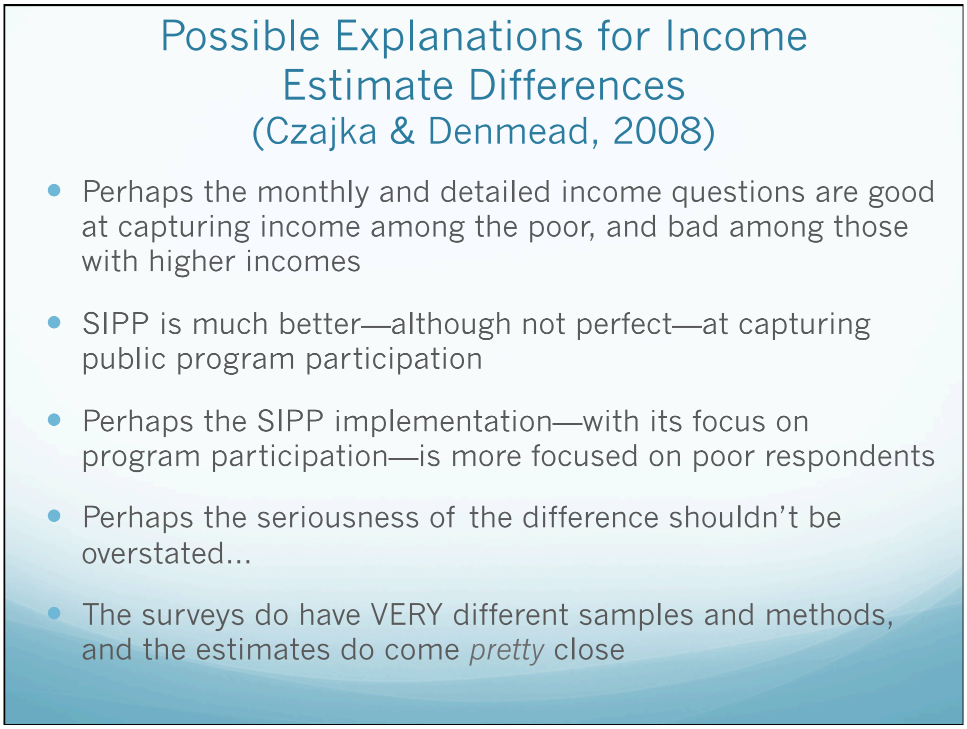
| Estimate | SIPP | CPS | ACS | MEPS |
|---|---|---|---|---|
| **Ave Family Income, Per Capita** | | | | |
| | $20,514 | 22,893 | 22,854 | 22,089 |
| **Family Income Per Capita by Quintile** | | | | |
| Lowest | $6,962 | 6,513 | 6,526 | 6,352 |
| Highest | $41,062 | 49,316 | 48,543 | 43,855 |



## Population Estimates By Survey

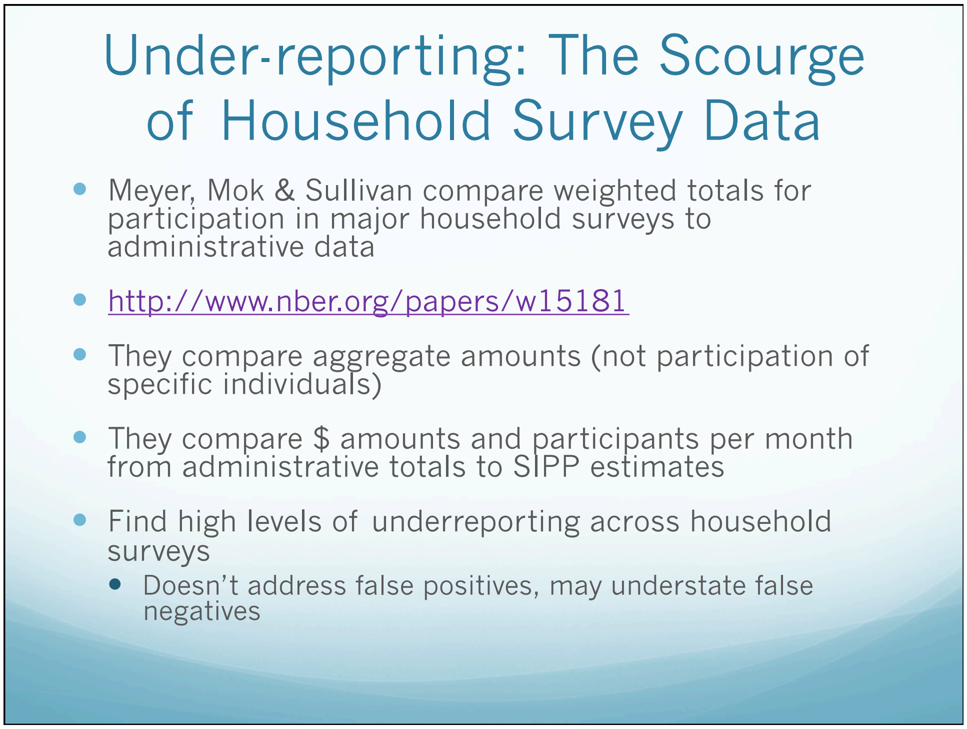
Calendar Year 2002 (Czajka & Denmead, 2008)

| Estimate (in millions) | SIPP | CPS | ACS |
|---|---|---|---|
| Total Population | 281 | 283 | 278 |
| < 100% Poverty | 33 | 34 | 35 |
| <200% Poverty | 56 | 52 | 49 |
| Children <100% Poverty | 13 | 12 | 13 |
| Receiving TANF/ SNAP | 31 | 21 | 24 |



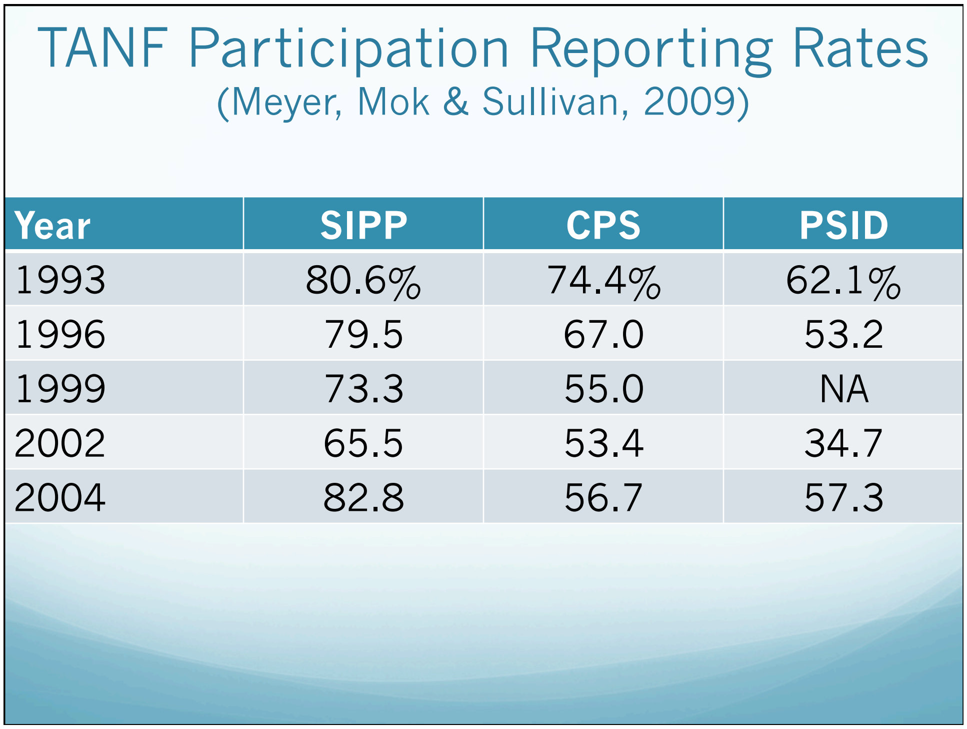
## Possible Explanations for Income Estimate Differences (Czajka & Denmead, 2008)

- Perhaps the monthly and detailed income questions are good at capturing income among the poor, and bad among those with higher incomes
- SIPP is much better—although not perfect—at capturing public program participation
- Perhaps the SIPP implementation—with its focus on program participation—is more focused on poor respondents
- Perhaps the seriousness of the difference shouldn't be overstated…
- The surveys do have VERY different samples and methods, and the estimates do come *pretty* close



## Under-reporting: The Scourge of Household Survey Data

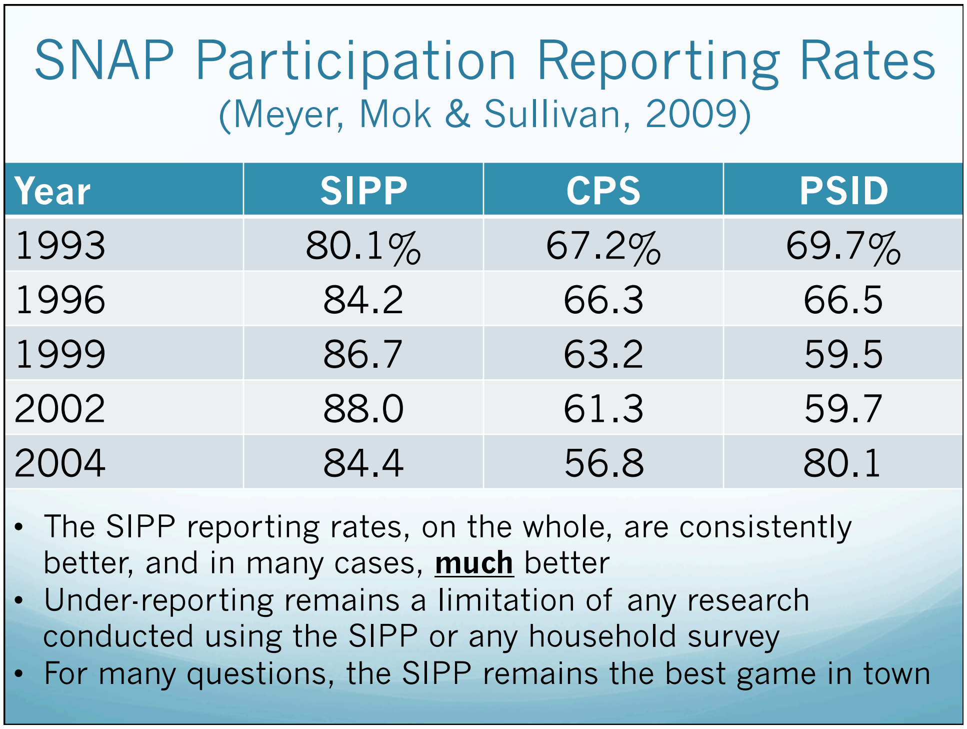
- Meyer, Mok & Sullivan compare weighted totals for participation in major household surveys to administrative data
- http://www.nber.org/papers/w15181
- They compare aggregate amounts (not participation of specific individuals)
- They compare $ amounts and participants per month from administrative totals to SIPP estimates
- Find high levels of underreporting across household surveys
  - Doesn't address false positives, may understate false negatives



## TANF Participation Reporting Rates

(Meyer, Mok & Sullivan, 2009)

| Year | SIPP | CPS | PSID |
|---|---|---|---|
| 1993 | 80.6% | 74.4% | 62.1% |
| 1996 | 79.5 | 67.0 | 53.2 |
| 1999 | 73.3 | 55.0 | NA |
| 2002 | 65.5 | 53.4 | 34.7 |
| 2004 | 82.8 | 56.7 | 57.3 |

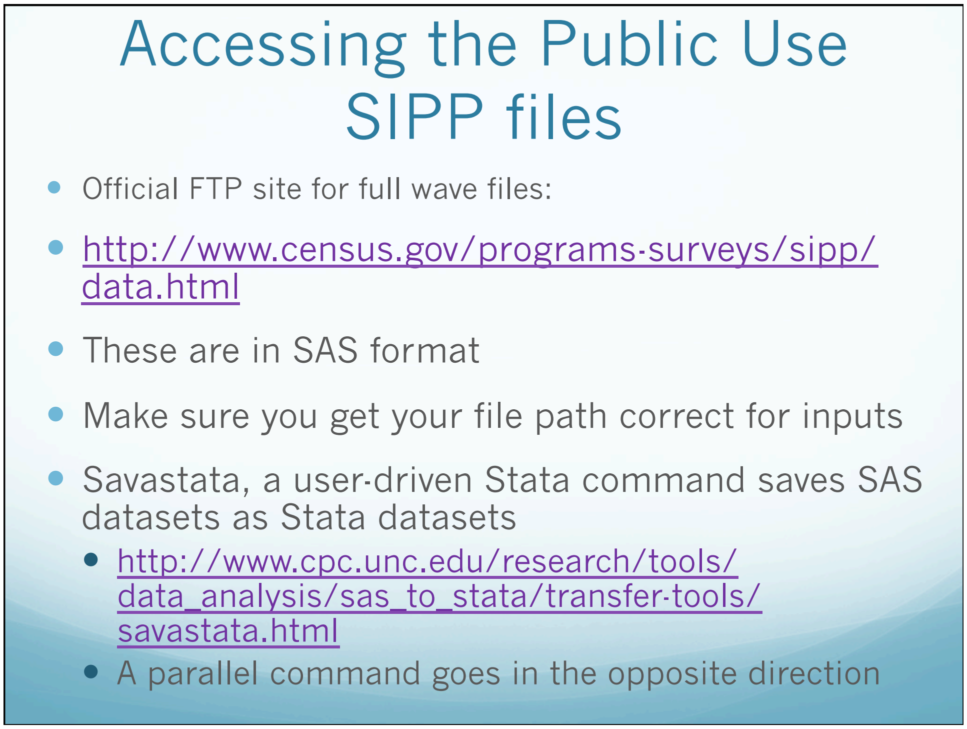


## SNAP Participation Reporting Rates

(Meyer, Mok & Sullivan, 2009)

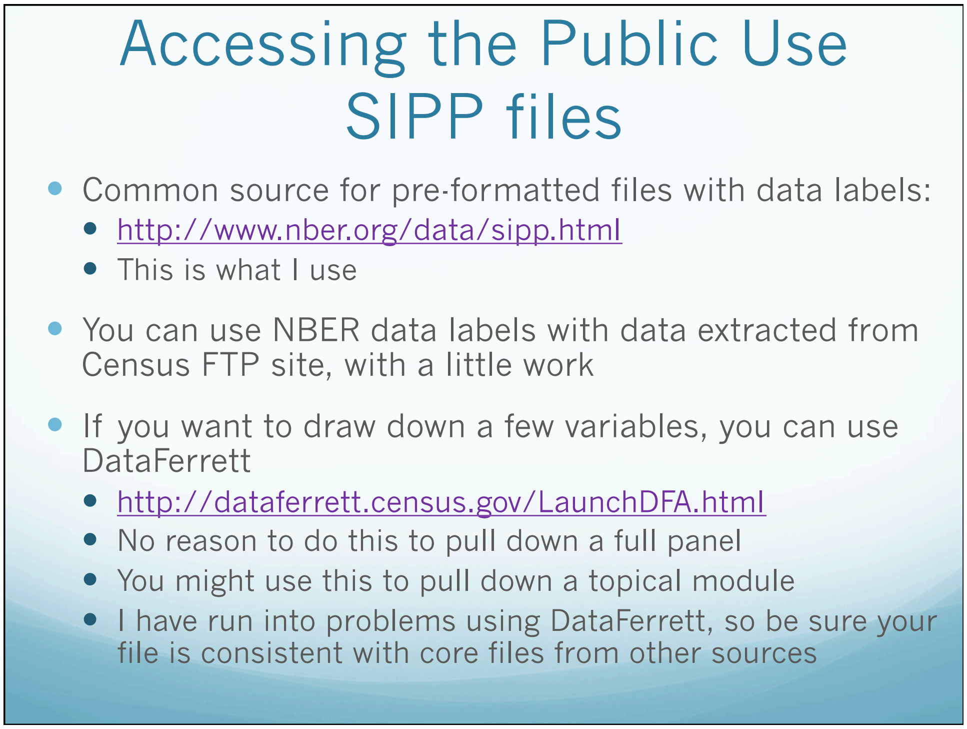
| Year | SIPP | CPS | PSID |
|---|---|---|---|
| 1993 | 80.1% | 67.2% | 69.7% |
| 1996 | 84.2 | 66.3 | 66.5 |
| 1999 | 86.7 | 63.2 | 59.5 |
| 2002 | 88.0 | 61.3 | 59.7 |
| 2004 | 84.4 | 56.8 | 80.1 |

- The SIPP reporting rates, on the whole, are consistently better, and in many cases, **much** better
- Under-reporting remains a limitation of any research conducted using the SIPP or any household survey
- For many questions, the SIPP remains the best game in town



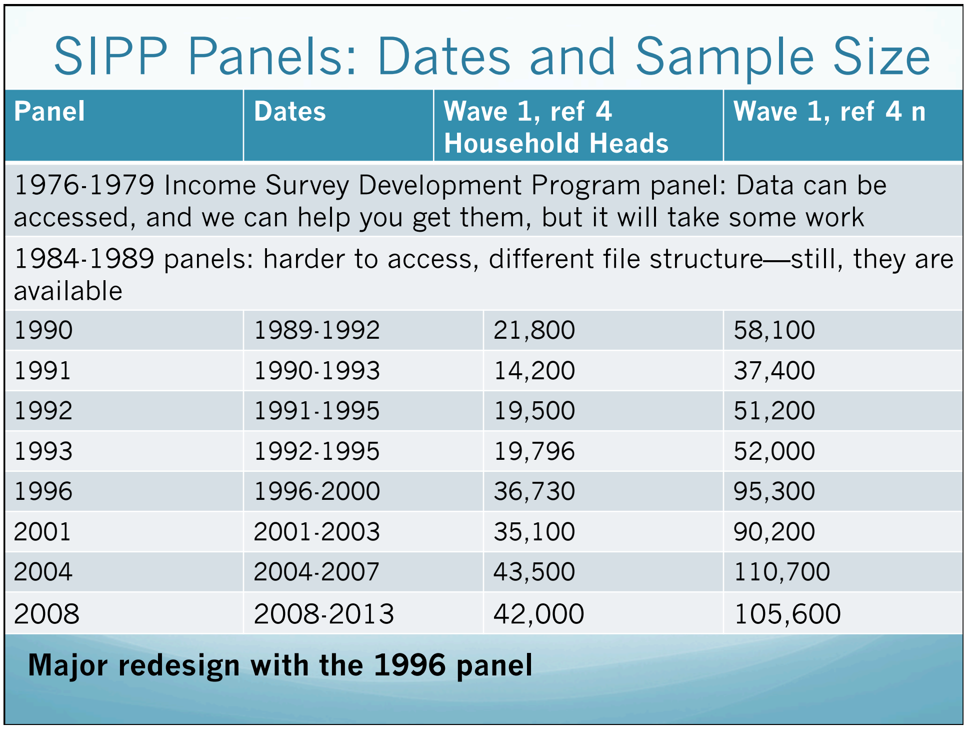
# Accessing the Public Use SIPP files

- Official FTP site for full wave files:
- http://www.census.gov/programs-surveys/sipp/data.html
- These are in SAS format
- Make sure you get your file path correct for inputs
- Savastata, a user-driven Stata command saves SAS datasets as Stata datasets
  - http://www.cpc.unc.edu/research/tools/data_analysis/sas_to_stata/transfer-tools/savastata.html
  - A parallel command goes in the opposite direction



# Accessing the Public Use SIPP files

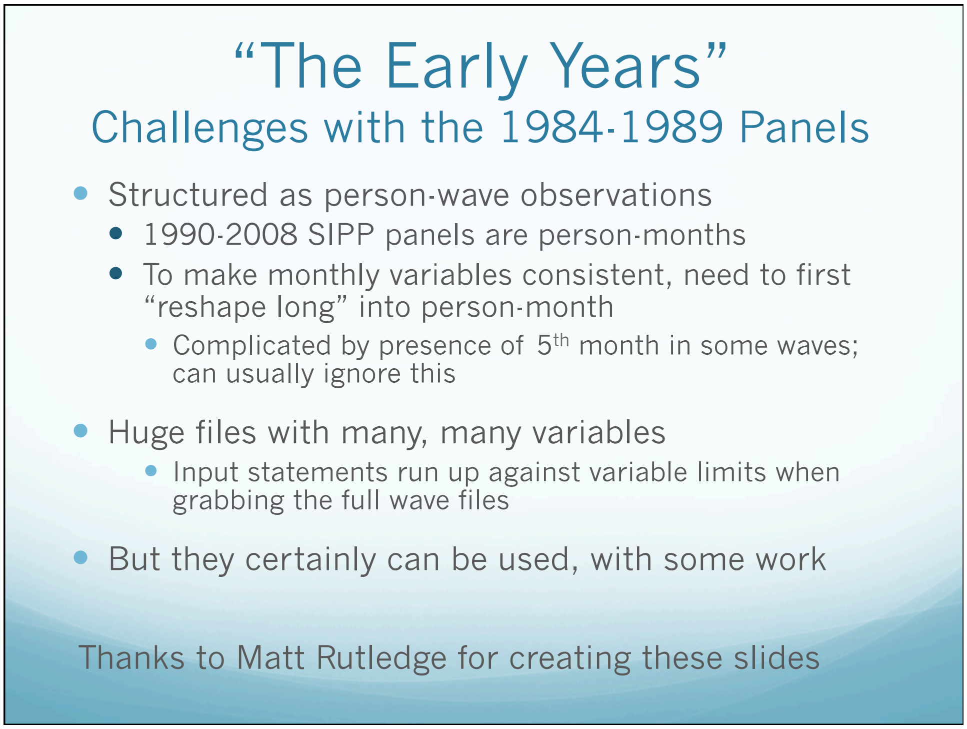
- Common source for pre-formatted files with data labels:
  - http://www.nber.org/data/sipp.html
  - This is what I use
- You can use NBER data labels with data extracted from Census FTP site, with a little work
- If you want to draw down a few variables, you can use DataFerrett
  - http://dataferrett.census.gov/LaunchDFA.html
  - No reason to do this to pull down a full panel
  - You might use this to pull down a topical module
  - I have run into problems using DataFerrett, so be sure your file is consistent with core files from other sources



# 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 can be accessed, and we can help you get them, but it will take some work | | | |
| 1984-1989 panels: harder to access, different file structure—still, they are available | | | |
| 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 redesign with the 1996 panel**

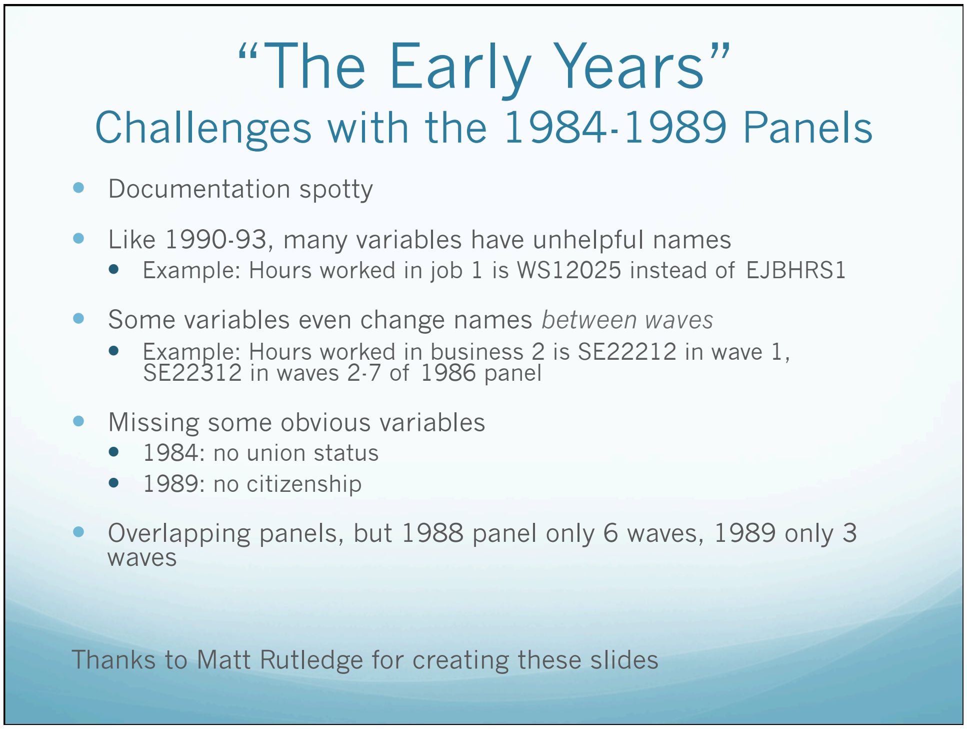


# "The Early Years"

## Challenges with the 1984-1989 Panels

- Structured as person-wave observations
  - 1990-2008 SIPP panels are person-months
  - To make monthly variables consistent, need to first "reshape long" into person-month
    - Complicated by presence of 5th month in some waves; can usually ignore this
- Huge files with many, many variables
  - Input statements run up against variable limits when grabbing the full wave files
- But they certainly can be used, with some work

Thanks to Matt Rutledge for creating these slides

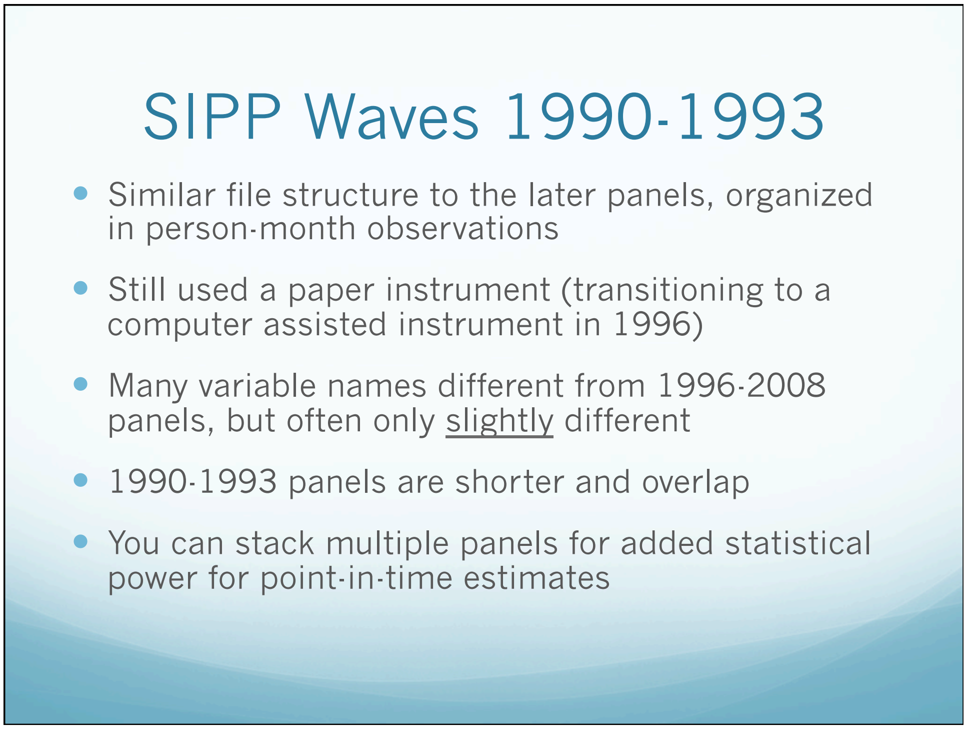


# “The Early Years”

## Challenges with the 1984-1989 Panels

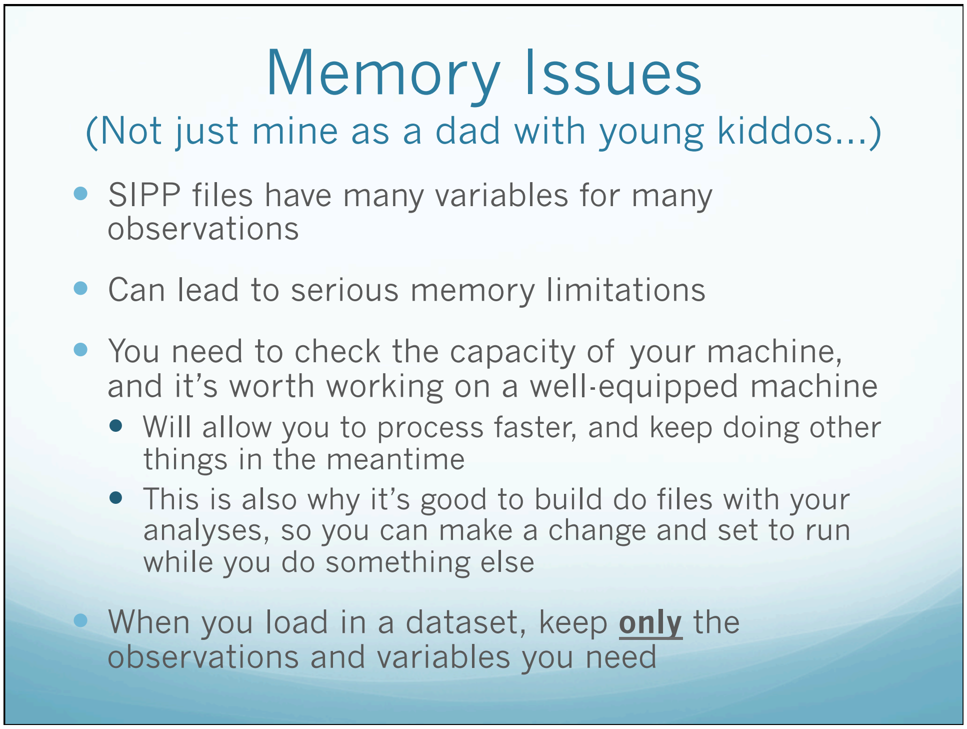
- Documentation spotty
- Like 1990-93, many variables have unhelpful names
  - Example: Hours worked in job 1 is WS12025 instead of EJBHRS1
- Some variables even change names *between waves*
  - Example: Hours worked in business 2 is SE22212 in wave 1, SE22312 in waves 2-7 of 1986 panel
- Missing some obvious variables
  - 1984: no union status
  - 1989: no citizenship
- Overlapping panels, but 1988 panel only 6 waves, 1989 only 3 waves

Thanks to Matt Rutledge for creating these slides



# SIPP Waves 1990-1993

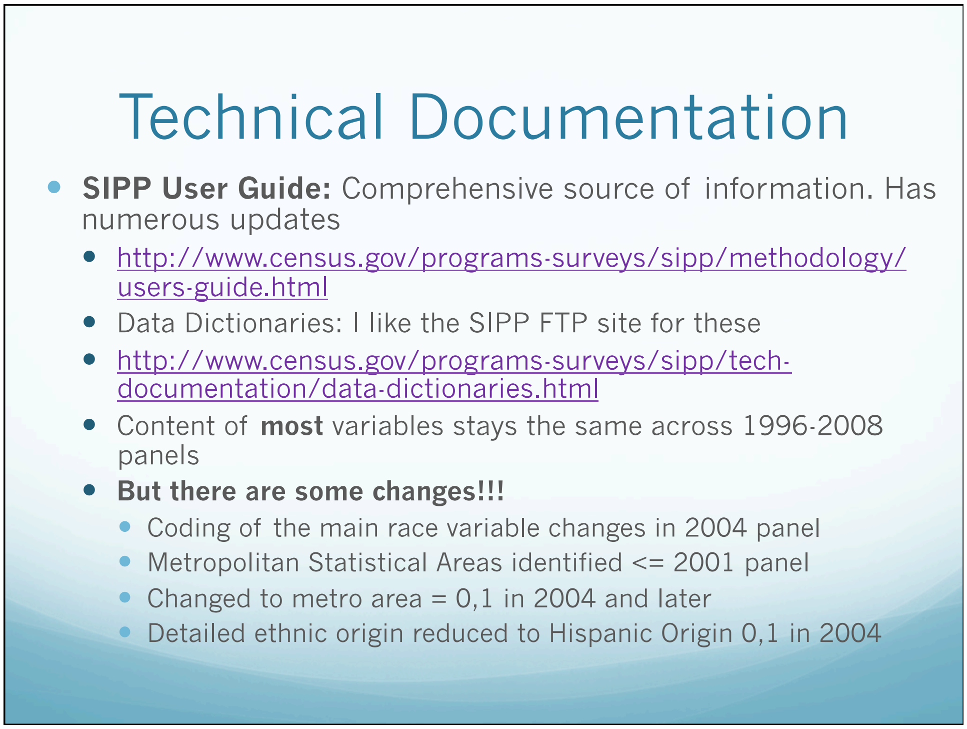
- Similar file structure to the later panels, organized in person-month observations
- Still used a paper instrument (transitioning to a computer assisted instrument in 1996)
- Many variable names different from 1996-2008 panels, but often only slightly different
- 1990-1993 panels are shorter and overlap
- You can stack multiple panels for added statistical power for point-in-time estimates



# Memory Issues

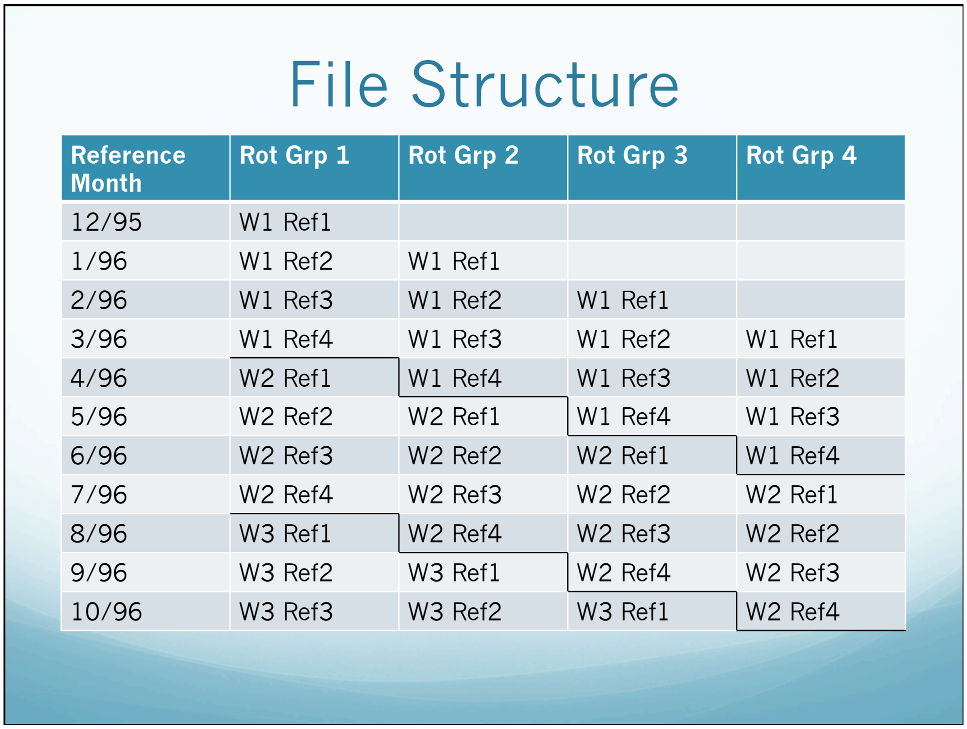
## (Not just mine as a dad with young kiddos…)

- SIPP files have many variables for many observations
- Can lead to serious memory limitations
- You need to check the capacity of your machine, and it's worth working on a well-equipped machine
  - Will allow you to process faster, and keep doing other things in the meantime
  - This is also why it's good to build do files with your analyses, so you can make a change and set to run while you do something else
- When you load in a dataset, keep **only** the observations and variables you need



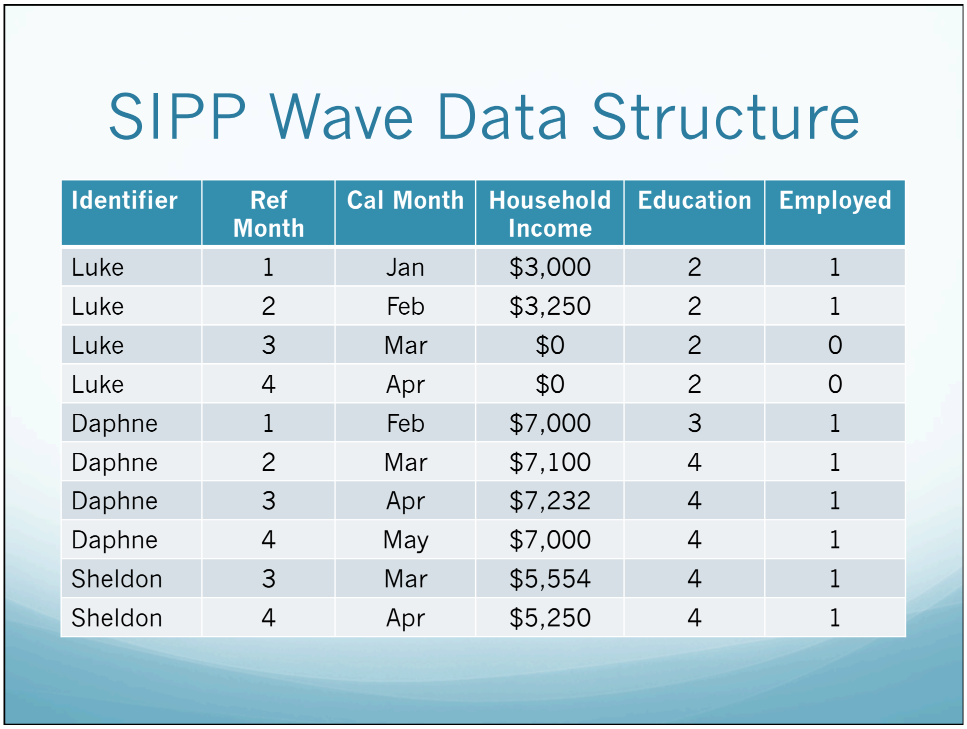
# Technical Documentation

- **SIPP User Guide:** Comprehensive source of information. Has numerous updates
  - http://www.census.gov/programs-surveys/sipp/methodology/users-guide.html
  - Data Dictionaries: I like the SIPP FTP site for these
  - http://www.census.gov/programs-surveys/sipp/tech-documentation/data-dictionaries.html
  - Content of **most** variables stays the same across 1996-2008 panels
  - **But there are some changes!!!**
    - Coding of the main race variable changes in 2004 panel
    - Metropolitan Statistical Areas identified <= 2001 panel
    - Changed to metro area = 0,1 in 2004 and later
    - Detailed ethnic origin reduced to Hispanic Origin 0,1 in 2004



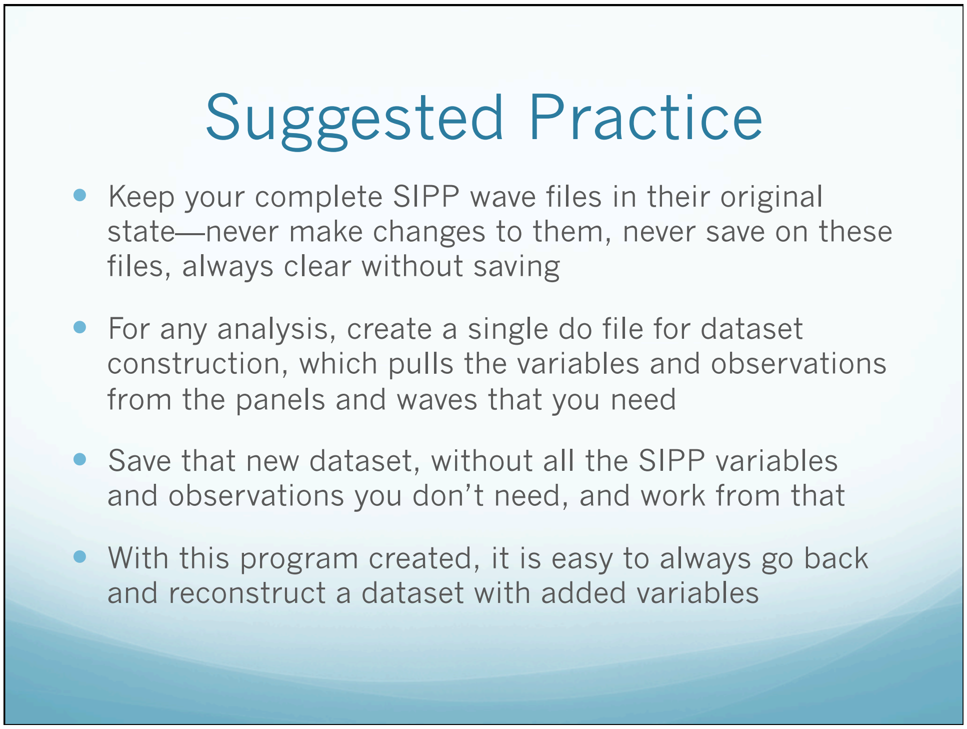
## File Structure

| 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 |



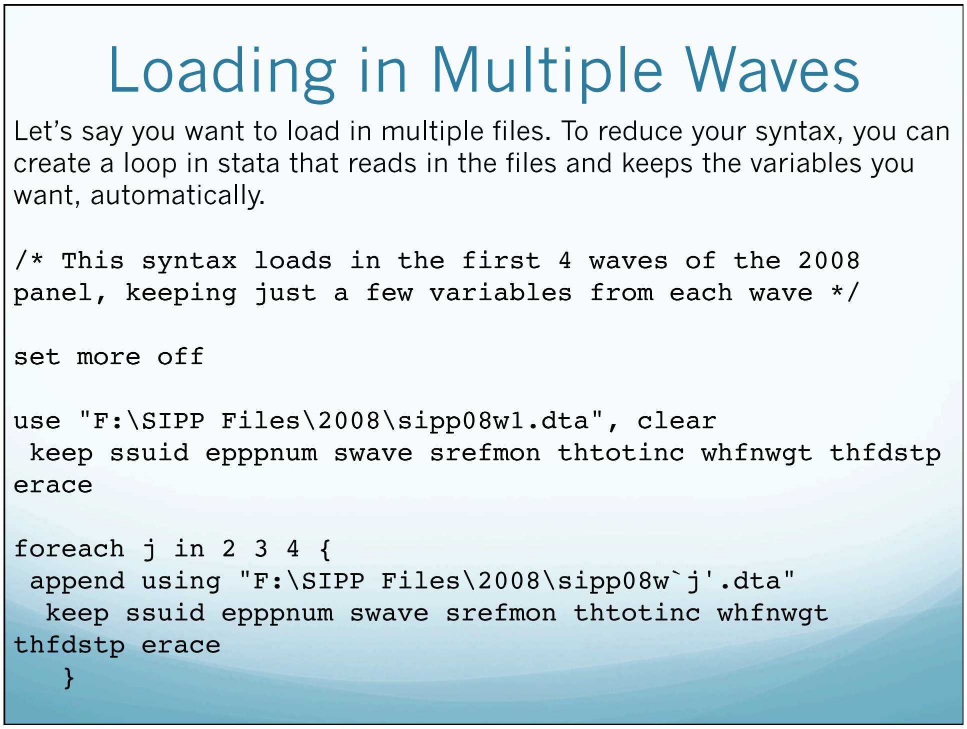
## SIPP Wave Data Structure

| Identifier | Ref Month | Cal Month | Household Income | Education | Employed |
|---|---|---|---|---|---|
| Luke | 1 | Jan | $3,000 | 2 | 1 |
| Luke | 2 | Feb | $3,250 | 2 | 1 |
| Luke | 3 | Mar | $0 | 2 | 0 |
| Luke | 4 | Apr | $0 | 2 | 0 |
| Daphne | 1 | Feb | $7,000 | 3 | 1 |
| Daphne | 2 | Mar | $7,100 | 4 | 1 |
| Daphne | 3 | Apr | $7,232 | 4 | 1 |
| Daphne | 4 | May | $7,000 | 4 | 1 |
| Sheldon | 3 | Mar | $5,554 | 4 | 1 |
| Sheldon | 4 | Apr | $5,250 | 4 | 1 |



# Suggested Practice

- Keep your complete SIPP wave files in their original state—never make changes to them, never save on these files, always clear without saving
- For any analysis, create a single do file for dataset construction, which pulls the variables and observations from the panels and waves that you need
- Save that new dataset, without all the SIPP variables and observations you don't need, and work from that
- With this program created, it is easy to always go back and reconstruct a dataset with added variables



# Loading in Multiple Waves

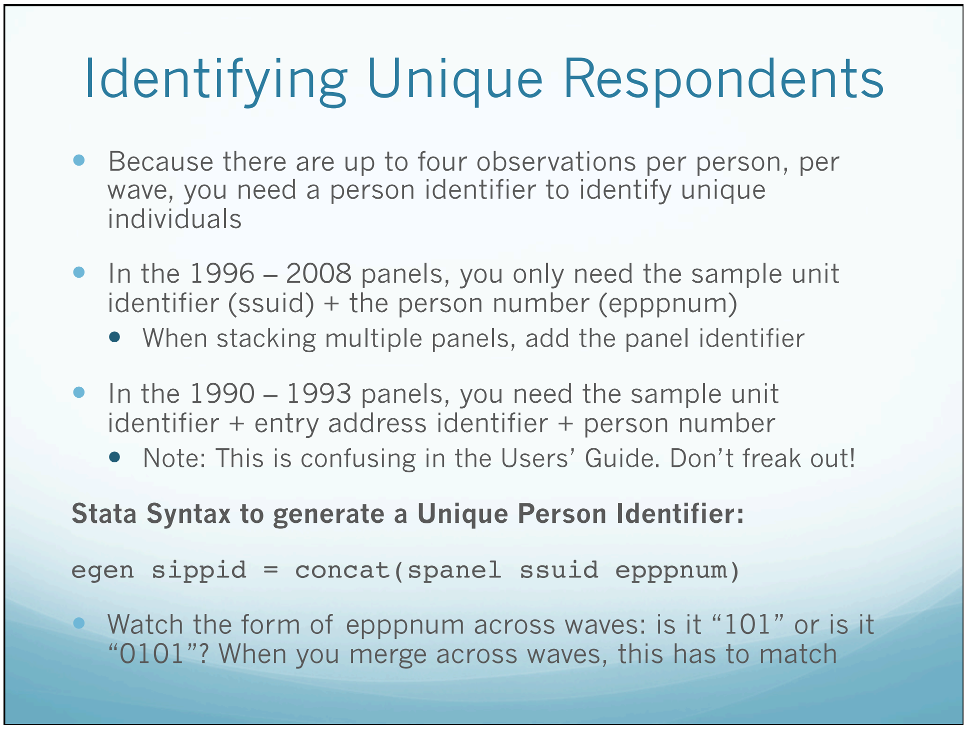
Let's say you want to load in multiple files. To reduce your syntax, you can create a loop in stata that reads in the files and keeps the variables you want, automatically.

```
/* This syntax loads in the first 4 waves of the 2008
panel, keeping just a few variables from each wave */

set more off

use "F:\SIPP Files\2008\sipp08w1.dta", clear
 keep ssuid epppnum swave srefmon thtotinc whfnwgt thfdstp
erace

foreach j in 2 3 4 {
 append using "F:\SIPP Files\2008\sipp08w`j'.dta"
  keep ssuid epppnum swave srefmon thtotinc whfnwgt
thfdstp erace
   }
```



# Identifying Unique Respondents

- Because there are up to four observations per person, per wave, you need a person identifier to identify unique individuals
- In the 1996 – 2008 panels, you only need the sample unit identifier (ssuid) + the person number (epppnum)
  - When stacking multiple panels, add the panel identifier
- In the 1990 – 1993 panels, you need the sample unit identifier + entry address identifier + person number
  - Note: This is confusing in the Users' Guide. Don't freak out!

**Stata Syntax to generate a Unique Person Identifier:**

```
egen sippid = concat(spanel ssuid epppnum)
```

- Watch the form of epppnum across waves: is it "101" or is it "0101"? When you merge across waves, this has to match