CPS Linking & Validation

Linking CPS Data

Opportunities and Challenges

Opportunities

- Large sample sizes
- Good coverage of subpopulations
- Short-run changes in employment and families; reactions to births and losses (death, divorce)
- Combine rich sources of information about different topics
  - Work and volunteering; veterans and employment

Same Month One Year Apart

- Useful for studying year-to-year change
  - Earnings and employment dynamics,
  - Geographic mobility,
  - Movement into and out of labor unions
- Researchers can expect:

	March 1987-1988	March 1994-1995	March 2009-2010
Linked N	54,890	48,131	53,486
Retention Rate	74.1%	69.3%	78.8%

Single Cohort, All 8 Months

- Useful for studying short-term dynamics
  - Change in economic arrangements as a function of social and demographic characteristics
  - Changes in these relationships over time
- Researchers can expect:

	Cohort Entering January 1987	Cohort Entering January 1994	Cohort Entering January 2009
Linked N	11,012	10,069	11,528
Retention Rate	59.0%	59.7%	68.0%

Challenges

- Rotation pattern
- Linking keys
- Same variables with different codes
- Non-response
- Another layer of complexity if ASEC is part of design
- Data management

Challenges: Rotation Pattern

- CPS is NOT a longitudinal survey that follows one or several cohorts as they age
- CPS does have a panel component where individuals are observed up to 8 times over 16 months, but the way they move through the survey is in a 4-8-4 rotation

Challenges: Rotation Pattern

- Enter IPUMS CPS RoPES
- Motivation
  - Easily *see* CPS rotation pattern
  - Explore what topical supplements can be examined together

Challenges

- Rotation pattern ✓
- Linking keys
- Same variables with different codes
- Non-response
- Another layer of complexity if ASEC is part of design
- Data management

Challenges: Linking Keys

- Changes in variables needed to link
  - Technical changes
  - Privacy protection changes
- Duplicate and recycled identifiers
  - Feng 2001
  - Flood et al. 2020

Challenges: Linking Keys

January 1976 to December 1993	January 1994 to April 2004	May 2004 +
HRHHID	HRHHID	HRHHID
PULINENO	PULINENO	PULINENO
STATEFIPS	STATEFIPS	STATEFIPS
HUHHNUM	HUHHNUM	
	HRSAMPLE	HRHHID2
	HRSERSUF	

**Transformations detailed in Drew, Flood & Warren 2014

Challenges: Linking Keys

- CPSID(P) – an IPUMS-created unique identifier
- Bridges changes in variable names as well as the technical aspects of how to link
- Available for 1976 forward
- Drew, Flood, & Warren, 2014 Journal of Economic and Social Measurement
- Flood et al., (forthcoming) Journal of Economic and Social Measurement

CPSID(P) Limitations

- Mechanical
  - Created solely based on linking keys
  - Does NOT condition on AGE, SEX, RACE
- Not available for all respondents in ASEC
  - only those who are also in the March BMS

CPSID(P) Limitations

- Unlinkable records
  - Records with duplicated identifiers in 1976-1983
  - Children are in some supplements but not basics in 1976-1981
- Unlinkable files
  - In 1976-1978, supplements and BMS-only months
  - Before and after July 1985
  - Before and after October 1985
  - Before and after June 1995

Future Linking Work

- Make ASEC oversamples linkable using CPSIDP
- A version of CPSIDP that is validated on AGE, SEX, RACE matches
  - Coming very soon!

Challenges

- Rotation pattern ✓
- Linking keys ✓
- Same variables with different codes
- Non-response
- Another layer of complexity if ASEC is part of design
- Data management

Challenges:
Same variables, different codes

- IPUMS!
- Harmonized across time to deal with changes in codes
- Use it for good, never for evil!
  - (Also, please cite us!)

Challenges

- Rotation pattern ✓
- Linking keys ✓
- Same variables with different codes ✓
- Non-response
- Another layer of complexity if ASEC is part of design
- Data management

Challenges: Non-response

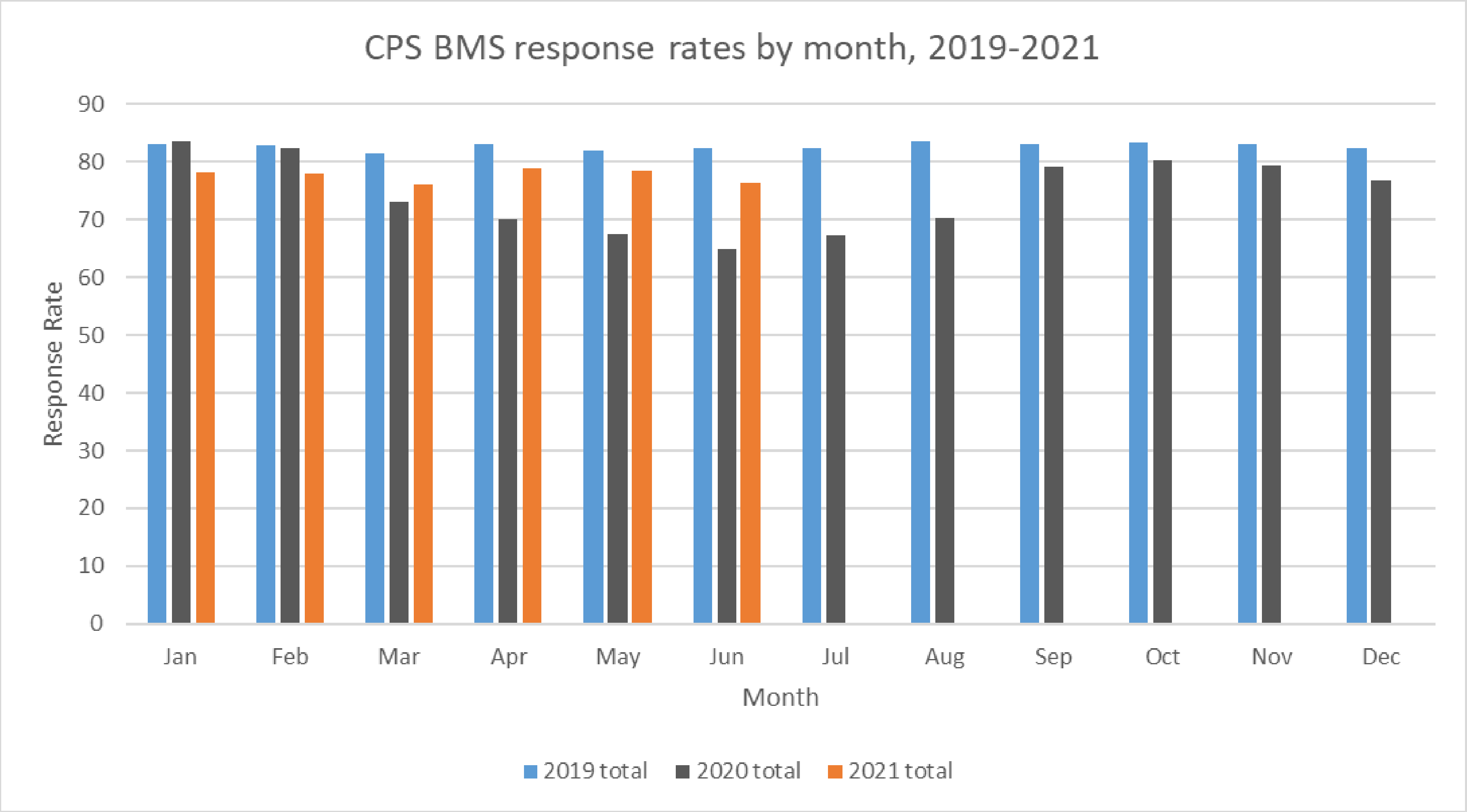
- Households that were eligible to participate in CPS but did not because of
  - Non-response (at any point in the rotation)
  - Death
  - Migration

- Imputation

Challenges: Non-response

- COVID-19
  - Response rates dropped noticeably during the pandemic, especially for incoming rotation groups
  - Individuals who link across months during the pandemic are more likely to be White, non-Hispanic, and to have higher income than those who are eligible to link, but do not (Rothbaum & Bee, 2020; Ward and Edwards, 2020)

Challenges: Non-response



Challenges

- Rotation pattern ✓
- Linking keys ✓
- Same variables with different codes ✓
- Non-response ✓
- Another layer of complexity if ASEC is part of design
- Data management

Challenges: ASEC

- Differently named variables, including linking keys, than basic monthly
- MARBASECID (1976, 1978+)
  - Variable to link ASEC and March basic monthly
  - Allows us to put CPSIDP on ASEC
  - Flood & Pacas 2017 Journal of Economic and Social Measurement
  - Flood et al., (forthcoming) Journal of Economic and Social Measurement

Challenges

- Rotation pattern ✓
- Linking keys ✓
- Same variables with different codes ✓
- Non-response ✓
- Another layer of complexity if ASEC is part of design ✓
- Data management

Challenges: Data Management

- Large number of files
- Needing to enforce rotation pattern to guard against ID recycling
- Strategies for dealing with this include:
  - Use IPUMS!
  - Loops
  - Temporary files
  - Long data format

Challenges

- Rotation pattern ✓
- Linking keys ✓
- Same variables with different codes ✓
- Non-response ✓
- Another layer of complexity if ASEC is part of design ✓
- Data management ✓

Validation

Approaches, Our Assumptions, and a Note on Data Structure

Validation: Approaches

- Because CPSIDP is mechanical, this is a step we recommend you do if you are linking CPS data across months
- No right or wrong way to do this, but there are a couple approaches in the literature
  1. Use AGE, SEX, RACE (Madrian & Lefgren)
  2. Use Bayesian approach (Feng)
     - Argues that some people are lost with first approach; this yields higher linkage rates

Validation: AGE, SEX, RACE Rules

- SEX & RACE
  - No change allowed
- AGE
  - If both time points are in MIS 1-4 or MIS 5-8, difference between AGE at two time points is 0 or 1
  - If the first time point is in MIS 1-4 and the second is in MIS 5-8, difference between AGE at two time points is 0, 1, or 2
  - Beware of AGE topcodes!
    - codes 80 and 85! Allow for a greater AGE increase. Age 80=80-84. Age 85=85+.

Validation: Age Rules

	MIS							
AGE	1	2	3	4	5	6	7	8
18								
50								
79								
84								

Can age 0 or 1 years compared to MIS 1

Can age 0 or 5 years compared to MIS 1

Can age 1 or 2 years compared to MIS 1

Can age 0 or 5 years compared to MIS 1

Validation: Our Assumptions

- For however many observations you are linking
  - SEX and RACE may not change
  - AGE may only change in expected ways
- Changes in AGE are always compared to the first time point being linked

Validation: Our Assumptions

- AGE, SEX, and RACE must ALL match in expected ways – GOOD!

MISH	AGE	SEX	RACE
Six	21	Female	White
Seven	21	Female	White
Eight	21	Female	White

Validation: Our Assumptions

- AGE, SEX, and RACE must ALL match in expected ways – BAD!

MISH	AGE	SEX	RACE
Six	21	Female	White
Seven	21	Female	Asian only
Eight	21	Female	White

Validation: For this Workshop

- Validation gets complicated quickly
- We want you to try to write validation code in the lab to think through validation
- We will provide validation code for you based on our all or nothing assumptions about AGE, SEX, and RACE for data in both
  - Long format
  - Wide format

Validation: Long Format

- For each individual
  - # records = # times observed

MISH	AGE	SEX	RACE
1	21	Female	White
2	21	Female	White
3	21	Female	White

Validation: Wide Format

- For each individual
  - # records = 1 regardless of # times observed

AGE			SEX			RACE		
1	2	3	1	2	3	1	2	3
21	21	21	Fem	Fem	Fem	White	White	White

Long to Wide Format Data

Long

MISH	AGE	SEX	RACE
1	21	Female	White
2	21	Female	White
3	21	Female	White

Wide

Questions?