SURV 622/SURVMETH 622: PARADATA

January 22, 2023

Fred Conrad

Definition

- Response data — Answers provided by respondents
- Metadata — Describes contents and context of data files, on survey level
- Paradata (Couper 1998) – Data about the process by which the survey data were collected, on individual interviewer- or case-level

Paradata overview

- Often captured by interviewers or survey software
- Types and ease of collection depend on the survey mode
  - Mail
  - Telephone or In-Person surveys
  - Web surveys

Paradata examples in CATI or CAPI surveys

- Interviewer characteristics
  - Common to know interviewer's age, sex, ethnicity
  - Could obtain other information such as personality traits or social skills
- Contact records
  - Time of each call/visit and outcome of each call/visit (using disposition code)
- Audit trails from interviewing software
  - Depending on software system, may capture the length of interview, clicking on help keys, comments entered
  - Some systems capture every keystroke

Paradata examples cont.

- Audio recordings
  - Could assess vocal properties of interviewer and respondent (e.g., pitch, disfluencies)
  - Could transcribe and code interviewer-respondent interaction (e.g., Was the interviewer interrupted while reading the question? Did interviewer read questions exactly as written?)
- Interviewer observations about the contact or interview
  - What does the interviewer believe the sex, age and race/ethnicity of the contact or respondent to be?
  - What were the stated reasons for refusal?

Paradata examples cont.

- Interviewer observations specific to CAPI surveys
  - Neighborhood – mix of business and residential units; upkeep of the yards or buildings; presence of litter or graffiti
  - Housing unit – whether sampled unit is a single-family home or in a multiunit structure; whether unit in a locked building or gated community; condition of unit relative to others in the area
  - Members of housing unit –presence of children under age 16; presence of non-English speakers

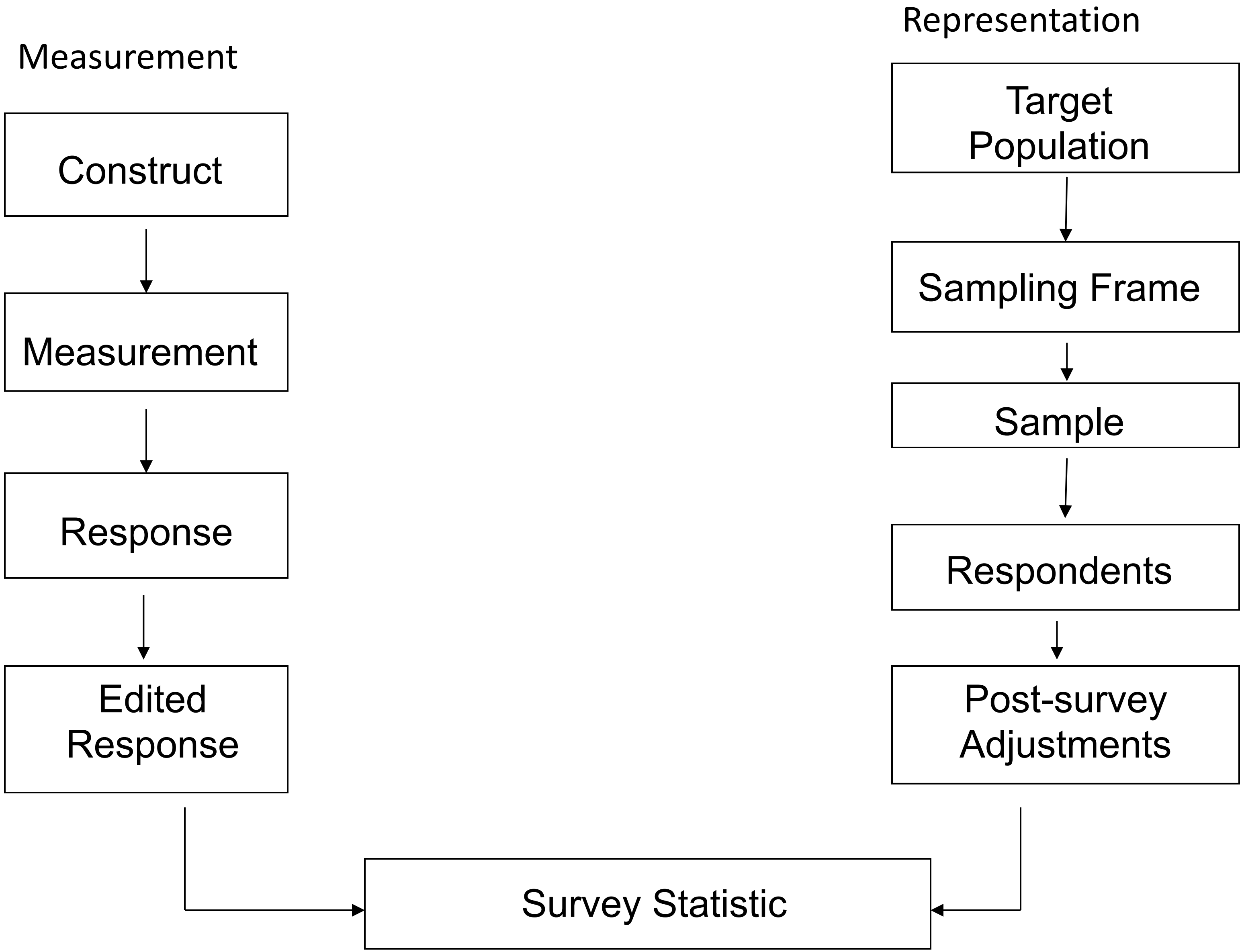
Paradata examples in web surveys

- Device type (smartphone, tablet, desktop)
- Response times (item- or page-level)
- Navigation (e.g., backups) and response changes
- Number of appearances of prompts or error messages

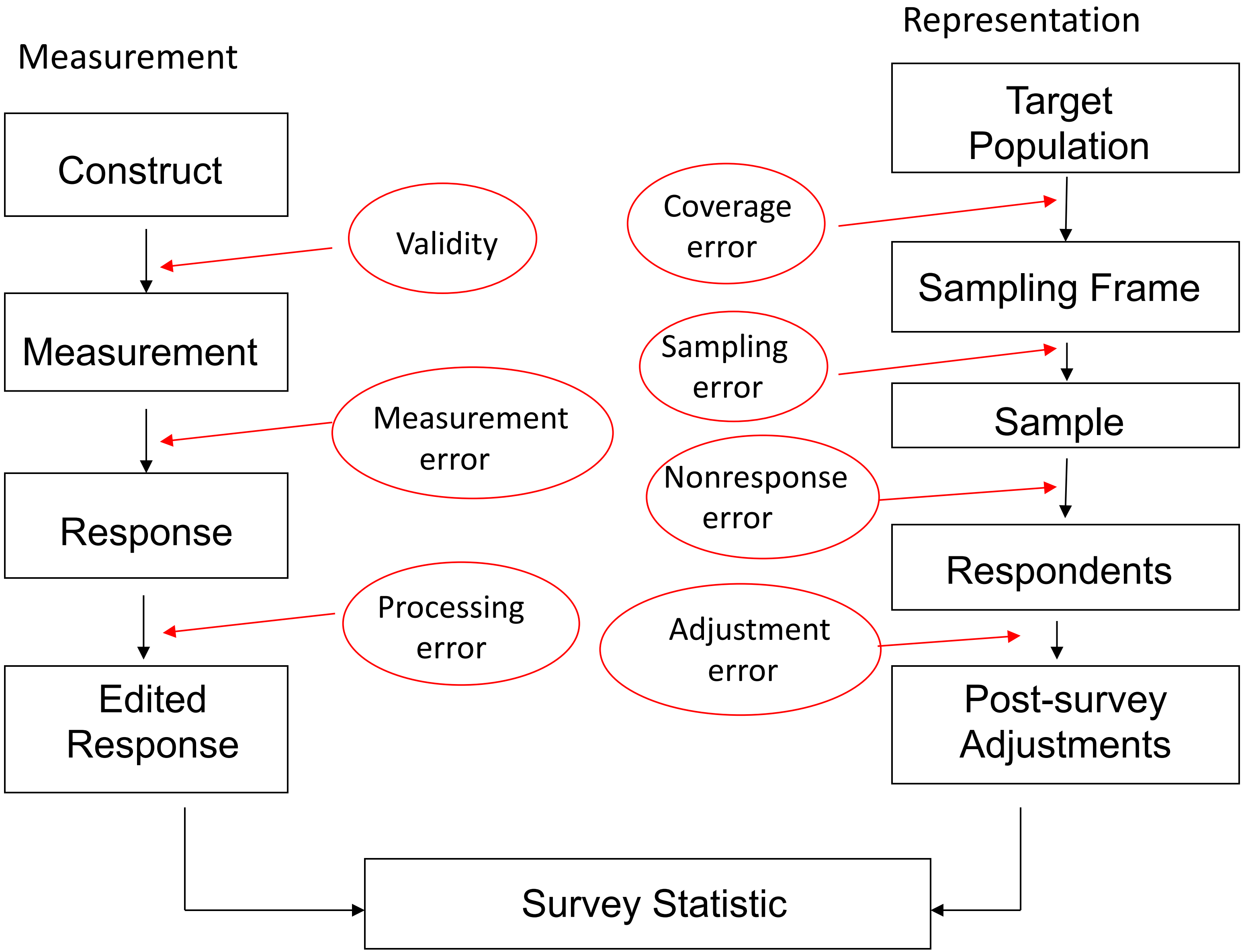
Potential uses for paradata

- Reducing nonresponse error
  - <u>Improving response rates </u>– e.g., data on contact rates by time of day may help to optimize calling patterns
  - <u>Informing “responsive design” strategies </u>— e.g., interviewer observations provide information about which cases are underrepresented; special efforts made to recruit those cases
  - <u>Improving nonresponse adjustments </u>– e.g., interviewer observations may provide information about non-respondents that can be used when creating nonresponse weights
- Reducing measurement error

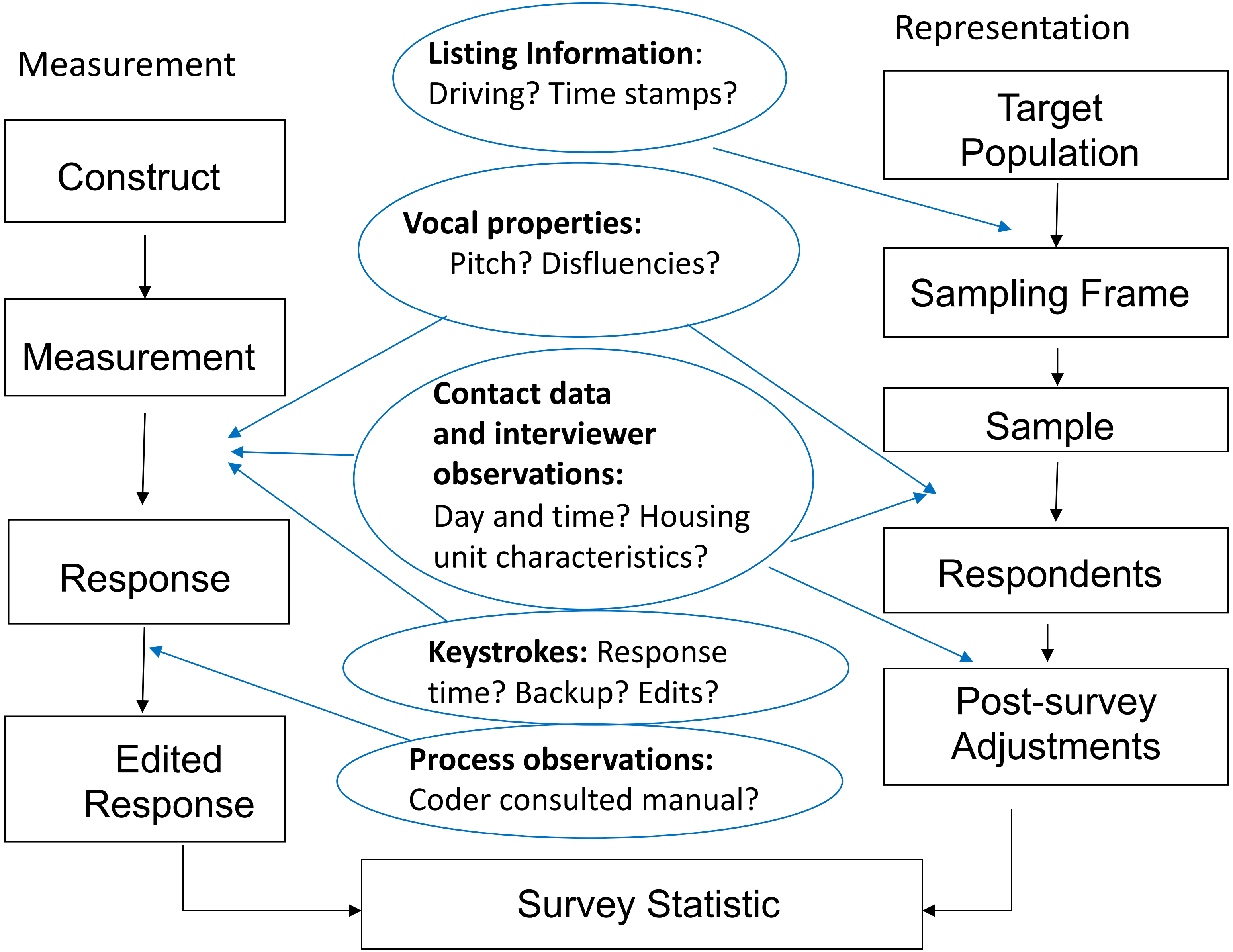
The survey life cycle



The survey life cycle and survey errors



The survey life cycle and survey paradata



Measurement

Construct

Measurement

Response

Edited Response

Representation

Target Population

Sampling Frame

Sample

Respondents

Post-survey Adjustments

Survey Statistic

Listing Information: Driving? Time stamps?

Vocal properties: Pitch? Disfluencies?

Contact data and interviewer observations: Day and time? Housing unit characteristics?

Keystrokes: Response time? Backup? Edits?

Process observations: Coder consulted manual?

Source: Adapted from Kreuter and Casas-Corderas (2010)

Using paradata to improve response rates

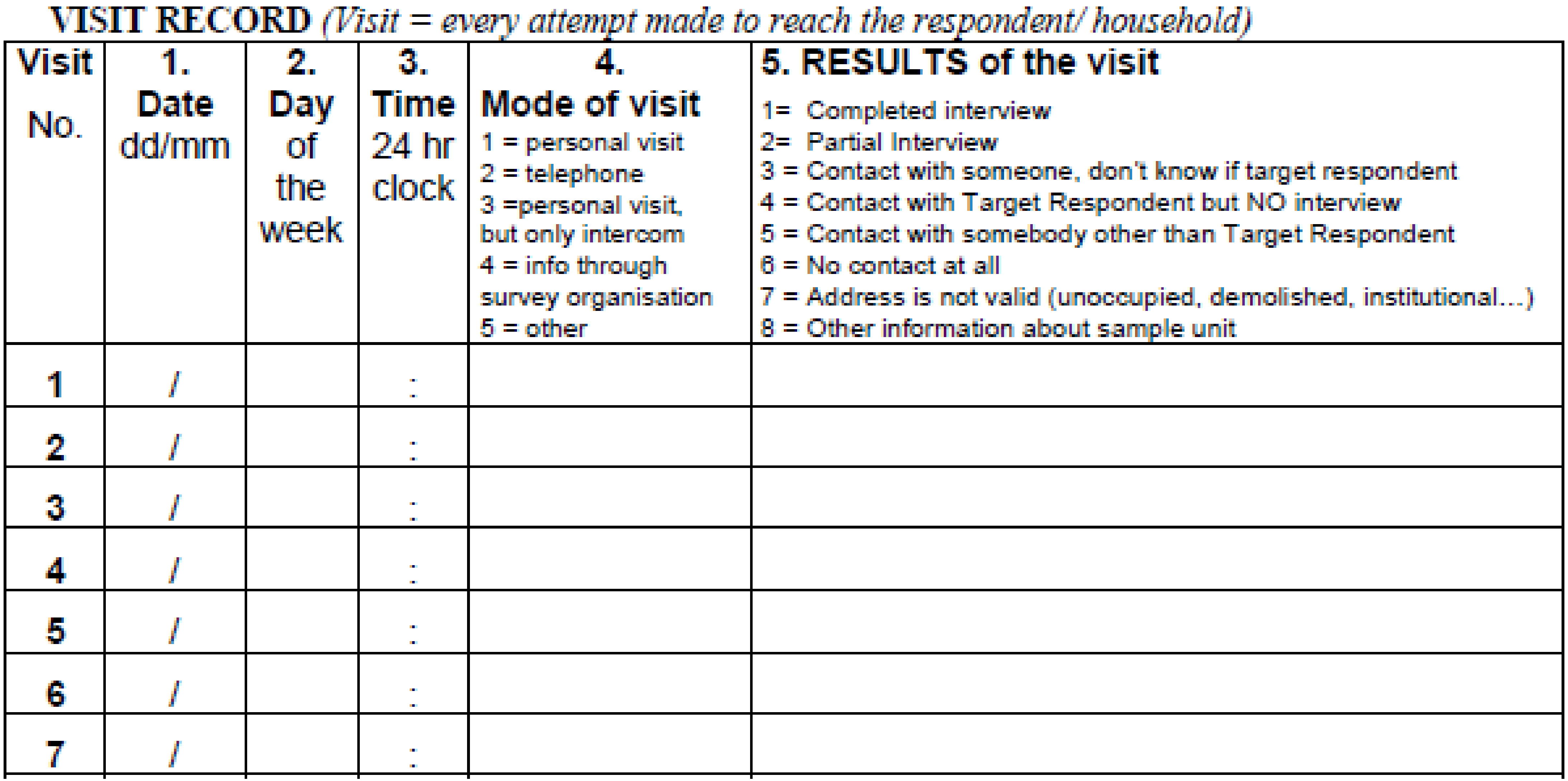
Information contained in call histories: CATI example

Case ID	Call ID	Date	Time	Outcome
10011	1	6/1/2012	3:12 PM	3130
10011	2	6/2/2012	10:34 AM	2111
10011	3	6/5/2012	6:23 PM	1000
10012	1	6/2/2012	11:42 AM	3140
10012	2	6/6/2012	4:31 PM	4700
10013	1	6/1/2012	9:31 AM	4510
10014	1	6/2/2012	10:04 AM	3130
10014	2	6/4/2012	9:42 AM	3130
10014	3	6/5/2012	7:07 PM	3130
10014	4	6/8/2012	5:11 PM	1000

Source: Kreuter and Olson (2013)

1000	Completed interview
2111	Refusal
3130	No answer
4510	Business, government office, other organization
4700	No eligible respondent

CAPI example

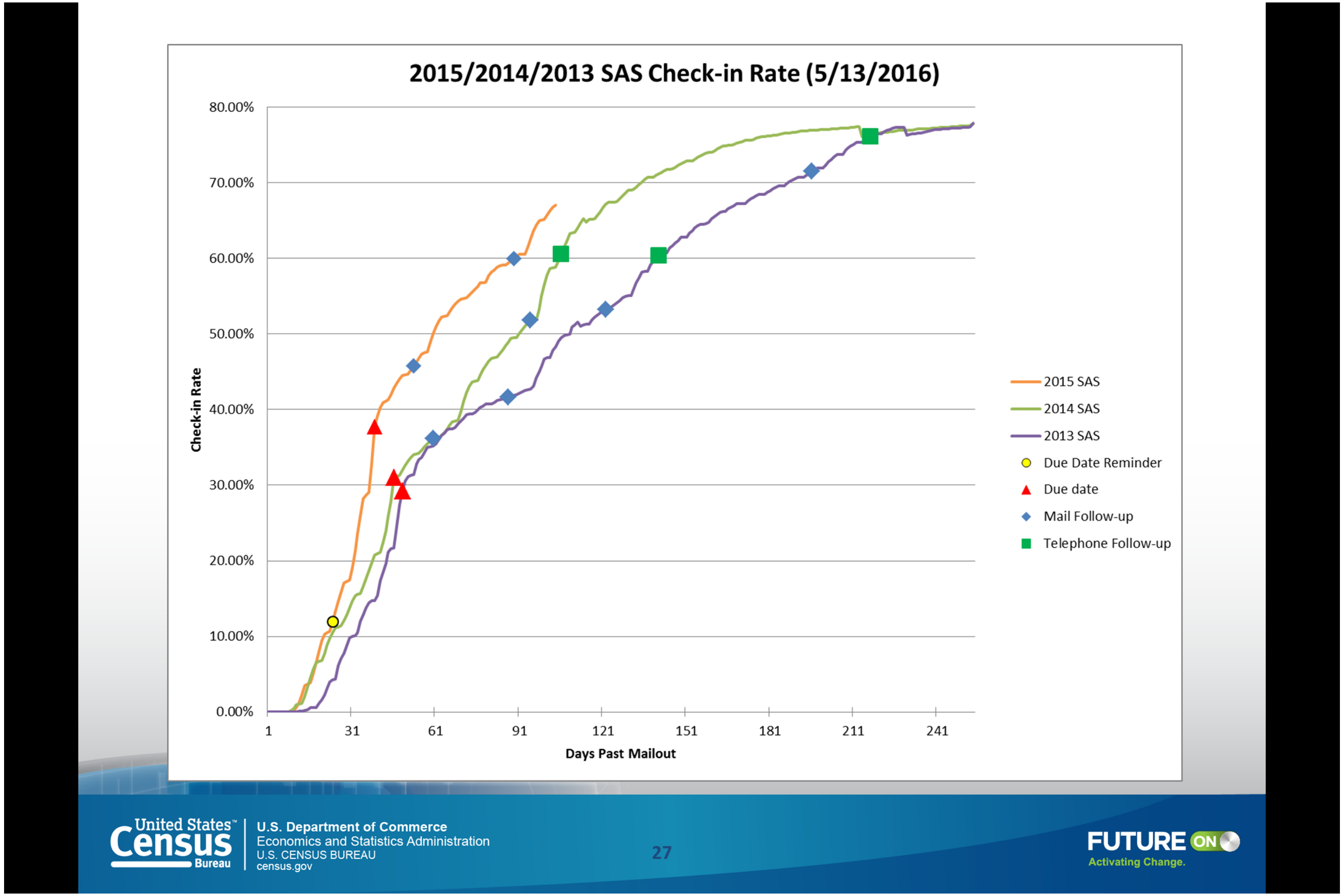


VISIT RECORD *(Visit = every attempt made to reach the respondent/ household)*

Visit No.	**1. Date** dd/mm	**2. Day** of the week	**3. Time** 24 hr clock	**4. Mode of visit** 1 = personal visit 2 = telephone 3 =personal visit, but only intercom 4 = info through survey organisation 5 = other	**5. RESULTS of the visit** 1= Completed interview 2= Partial Interview 3 = Contact with someone, don't know if target respondent 4 = Contact with Target Respondent but NO interview 5 = Contact with somebody other than Target Respondent 6 = No contact at all 7 = Address is not valid (unoccupied, demolished, institutional…) 8 = Other information about sample unit
1	/		:		
2	/		:		
3	/		:		
4	/		:		
5	/		:		
6	/		:		
7	/		:		

Source: European Social Survey Round 6 documentation

Survey dashboards



United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

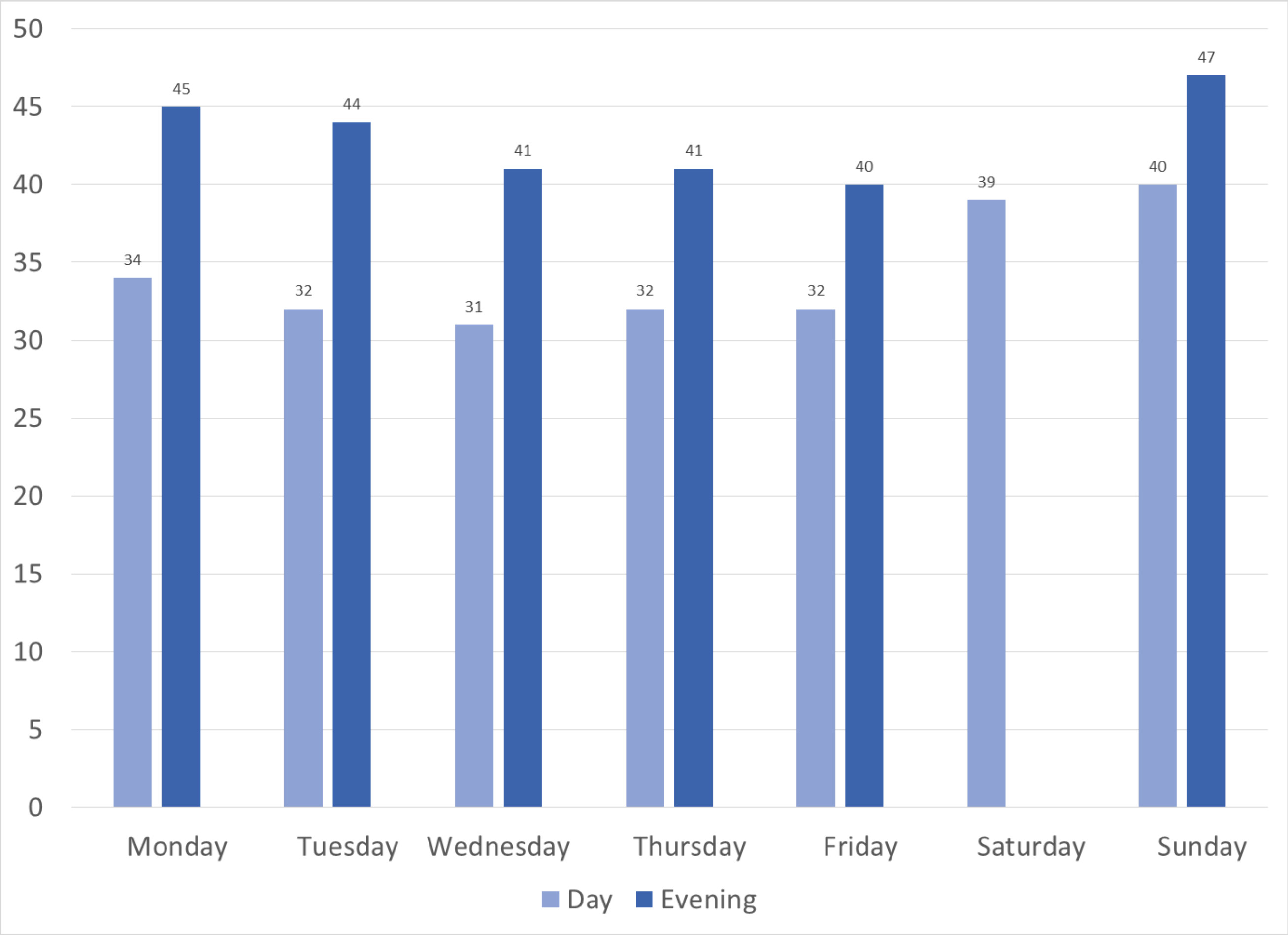
27

FUTURE ON
Activating Change.

Services Annual Survey

What do paradata say about best time to call? Analyses of call records have yielded some “rules of thumb”.

- Day of week and time of day matter (e.g., Laflamme 2008)
- Second or third call in same window as first call less likely to be productive (e.g., Kulka and Weeks 1988)
- In panel surveys, call household on day and at time contacted in previous wave (e.g., Laurie and Smith 1999)



Contact rates by day of week and time of day. Travel Activities and Motivation Survey, Statistics Canada. Source: Laflamme (2008)

Example: Using paradata to optimize contacts for individual survey cases (Wagner 2013)

- Experiments implemented in an RDD survey (Survey of Consumer Attitudes) and an in-person survey (National Survey of Family Growth)
  - Survey cases randomly assigned to experimental versus control condition
  - Control cases handled in the normal fashion
- Experimental protocol
  - Define call windows (day and time)
  - Estimate probability of contact for each household-window combination
  - Within each window, prioritize cases for which highest probability of contact was within that window relative to other windows
  - Place calls based on algorithm (telephone) or make recommendations to interviewers about when to attempt contact (face to face)
  - Re-estimate probabilities
  - Repeat until fielding period ends
- Outcome of interest the contact rate under two conditions

Wagner (2013) cont.

- Initial models for probability of contact during given survey window based on data from previous survey waves
  - Neighborhood context variables at the Census block level, numbers in exchange that are listed (RDD)
- After first round of contacts, information on experience with individual survey cases added to the model
  - Over time, individual information plays a bigger role in determining estimated probabilities

Wagner (2013) cont.

Illustration: estimated contact probabilities by time window

Case	Window 1	Window 2	Window 3	Window 4
1	0.05	0.01	0.03	0.02
2	0.25	0.35	0.20	0.15
3	0.05	0.10	0.15	0.08
4	0.40	0.50	0.30	0.20

– Priority ordering of cases in Window 1?

– Priority ordering of cases in Window 2?

Wagner (2013) cont.

- Results in this case show little improvement in contact rates compared to standard protocol
- Why? Some possible contributing factors:
  - Algorithm may not be optimal
    - Especially difficult to determine best time to call refusal conversion cases
  - In CAPI survey, errors in visit records may have biased models
  - In CAPI survey*, Iwers* did not appear to follow recommendations they were given – visits are scheduled in groups, so interviewers do not see recommendations for individual housing units as useful
  - Existing protocols have been refined over time so that they are reasonably good

Paradata and nonresponse adjustment

Nonresponse reweighting adjustment

- Nonresponse bias depends on two things:
  - Level of nonresponse
  - Differences between respondents and nonrespondents:

$$B(\overline{y}_r) = \overline{y}_r - \overline{y}_t = (\frac{n_{nr}}{n})(\overline{y}_r - \overline{y}_{nr})$$

- Standard reweighting adjustments require variables that are:
  - Known for both respondents and nonrespondents
  - Correlated with the probability of response
  - Correlated with the variables of interest
- Seek paradata with these properties to augment information typically available on the survey frame

Hypothetical example: Using interviewer observations for nonresponse reweighting

- Survey to assess views of an upcoming school bond issue
- Sample stratified by income level of Census block
  - Standard nonresponse weights based on neighborhood income quartile

Neighborhood income quartile	n	n_R	ω_{NR}
1	500	450	
2	500	400	
3	500	350	
4	500	300	

Example cont.

Neighborhood income quartile	n	n_R	ω_{NR}
1	500	450	1.11
2	500	400	1.25
3	500	350	1.43
4	500	300	1.67

Example cont.

- Suppose interviewer observations on presence of children can be obtained for all sample households (not just respondents)
    - Can be used to construct a more refined set of nonresponse adjustment weights

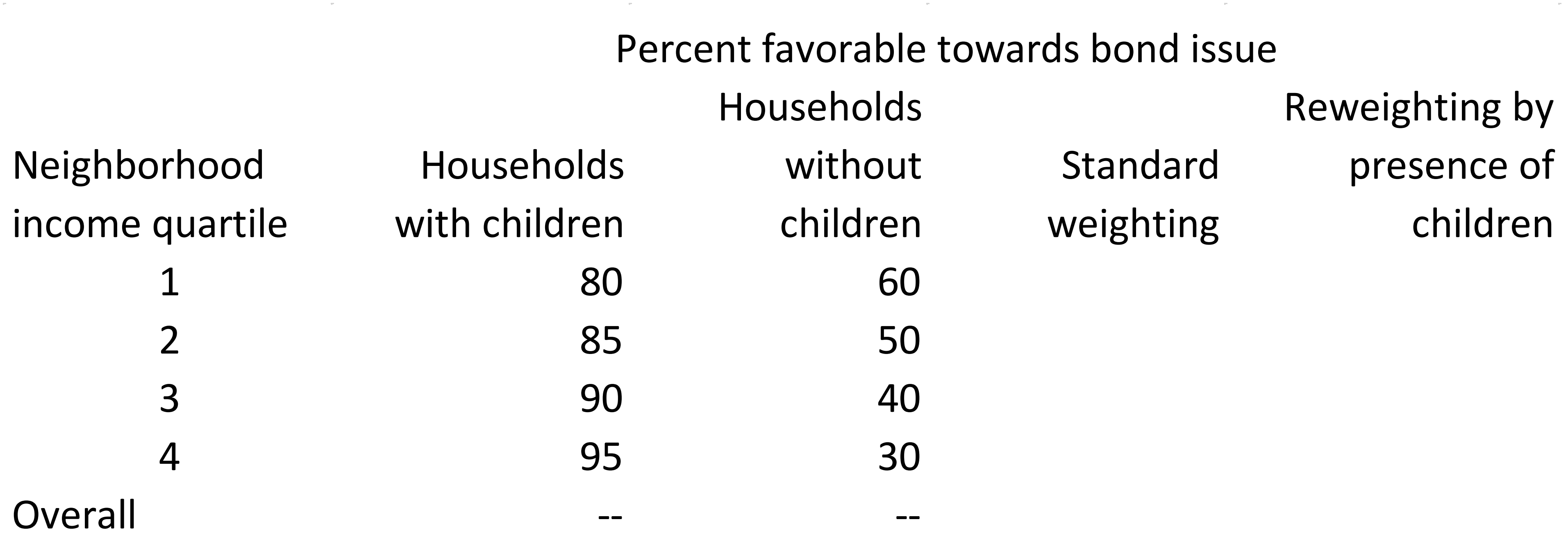
Neighborhood income quartile	n^{C}	n^{NC}	n^{C}_{R}	n^{NC}_{R}	ω^{C}_{NR}	ω^{NC}_{NR}
1	400	100	400	50		
2	350	150	350	50		
3	300	200	300	50		
4	250	250	250	50		

Example cont.

Neighborhood income quartile	n^{C}	n^{NC}	n^{C}_{R}	n^{NC}_{R}	ω^{C}_{NR}	ω^{NC}_{NR}
1	400	100	400	50	1.00	2.00
2	350	150	350	50	1.00	3.00
3	300	200	300	50	1.00	4.00
4	250	250	250	50	1.00	5.00

Example cont.

- Suppose further that respondents with and without children have different views about proposed bond issue
- Under these circumstances, weights informed by paradata may improve the estimates



	Percent favorable towards bond issue			
		Households		Reweighting by
Neighborhood	Households	without	Standard	presence of
income quartile	with children	children	weighting	children
1	80	60		
2	85	50		
3	90	40		
4	95	30		
Overall	--	--		

Example cont.

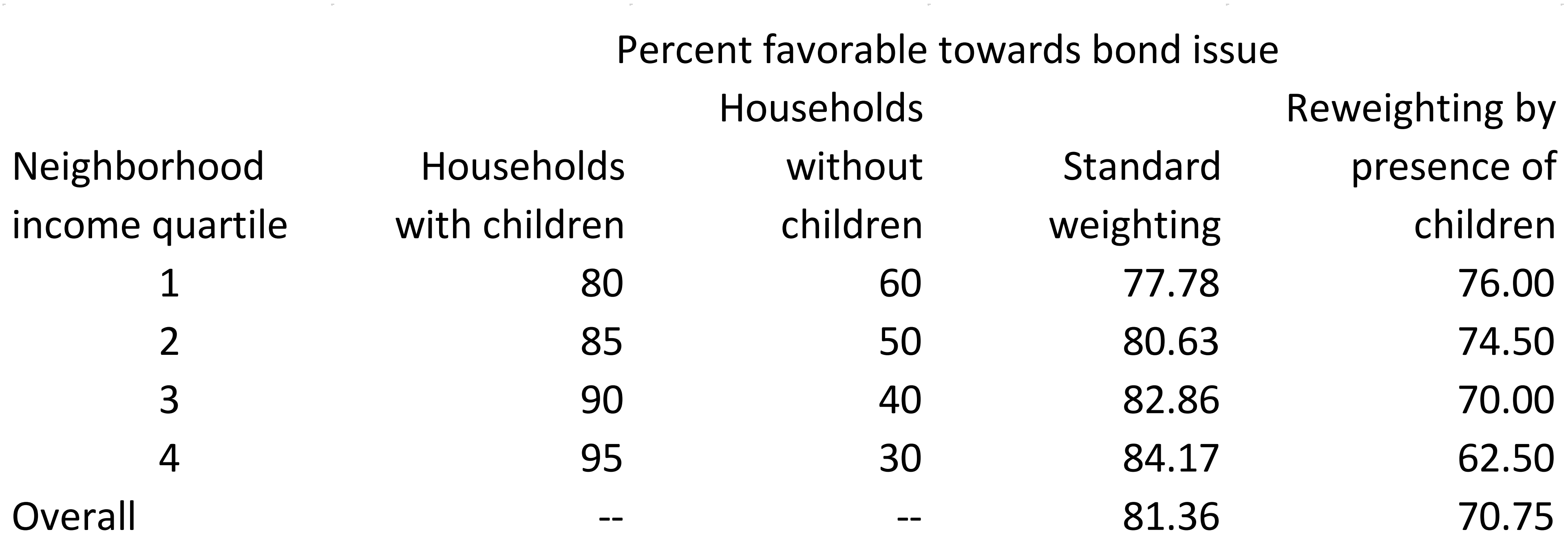
- To illustrate, estimates in <u>first row of table</u> in previous slide would be calculated as follows

  Standard weighting:
  [(400*80*1.11)+(50*60*1.11)]/500 = 77.78

  Reweighting by presence of children:
  [(400*80*1.00)+(50*60*2.00)]/500 = 76.00

- Can you fill in the other missing numbers?

Example cont.



	Percent favorable towards bond issue			
		Households		Reweighting by
Neighborhood	Households	without	Standard	presence of
income quartile	with children	children	weighting	children
1	80	60	77.78	76.00
2	85	50	80.63	74.50
3	90	40	82.86	70.00
4	95	30	84.17	62.50
Overall	--	--	81.36	70.75

Example cont.

- Suppose that, within each neighborhood income group, households with and without children had the same response rate (i.e., no relationship between paradata measure and probability of response). What happens?

Neighborhood income quartile	n^{C}	n^{NC}	n^{C}_{R}	n^{NC}_{R}	ω^{C}_{NR}	ω^{NC}_{NR}
1	400	100	360	90	1.11	1.11
2	350	150	280	120	1.25	1.25
3	300	200	210	140	1.43	1.43
4	250	250	150	150	1.67	1.67

Example cont.

- In this example, the estimates in first neighborhood would be calculated as follows

Standard weighting:
[(360*80*1.11)+(90*60*1.11)]/500 = 75.92

Reweighting by presence of children:
[(360*80*1.11)+(90*60*1.11)]/500 = 75.92

- Reweighting by presence of children has no effect

Example cont.

- Now suppose that, within each neighborhood income group, households with and without children have same views about bond issue (i.e., no relationship between paradata measure and variable of interest). What happens?

Neighborhood income quartile	Households with children	without children
1	80	80
2	85	85
3	90	90
4	95	95

Example cont.

- In this example, the estimates in first neighborhood would be calculated as follows

  Standard weighting:
  [(400*80*1.11)+(50*80*1.11)]/500 = 80.00

  Reweighting by presence of children:
  [(400*80*1.00)+(50*80*2.00)]/500 = 80.00

- Reweighting by presence of children has no effect

Challenges of using paradata

- In practice, easier to identify observational paradata that are predictive of response than observational paradata that are correlated with variable(s) of interest
  - E.g., number of call attempts may only be correlated with participation
- Interviewer judgments may be incomplete or erroneous
  - Based on experiments done in the General Social Survey in 1996 and 2000, Saperstein (2006) reports agreement of interviewer identification with self-reported race 99% for whites, 97% for blacks but 50% for other
  - Similar analysis found interviewer age observation placed household within correct 10-year age band 76% of time (Sinibaldi 2010)

Challenges cont.

- In a simulation study, West (2013) shows that even modest errors can substantially reduce the effectiveness of nonresponse adjustments that rest on paradata
  - Better interviewer training could reduce observation errors (including missing observations)
- Working with call histories, audit trails and audio recordings can be difficult and labor intensive
  - Methodologies still being developed
- Privacy concerns may preclude release of (some) paradata

Summing up

- Rich array of paradata available for Telephone, In-person and Web surveys
- Paradata may help with addressing both errors of representation and errors of measurement
- Paradata dashboards becoming standard for management of survey response process
- Active research on collection and use of paradata to address errors of representation
  - Improvements in paradata quality
  - Increase response rates
  - Improve non-response adjustments