Analysis with Survey Data

Dr. Richard W. Evans

October 24, 29, 2018

Two methods for asking questions

1. Conducting survey
   - Large numbers of participants
   - Highly structured questionnaires
   - Statistical methods to create representativeness
   - More comparable across observations
2. Interviews
   - Small number of participants
   - Less structure, more organic questioning interaction
   - Rich qualitative information
   - Difficult to systematize

Focus on Survey

"...digital age creates many exciting opportunities for survey researchers to collect data more quickly and cheaply, to ask different kinds of questions, and to magnify the value of survey data with [large] data sources." (Salganik, 2018, p. 85)

Survey vs. Observational data

1. Digital observational data sometimes have problems with:
    - accuracy, completeness, accessibility, interpretability
2. Surveys can get at *internal states*
    - emotions, knowledge, expectations, opinions

Eras of survey research design

1. 1930s-1950s
   - Door-to-door, in-person questionnaire
   - Beginnings of representativeness
2. 1960s-1990s
   - Telephone surveys
   - Random digit dialing
   - Increased sampling efficiency from theory
3. 2000s-present
   - E-mail and internet surveys
   - More respondents
   - Bigger issues with bias and selection

Eras of survey research design

Salganik prediction

"I expect that the third era of survey research will be characterized by non-probability sampling, computer-administered interviews, and the linkage of surveys to [large] data sources." (Salganik, 2015, p. 86)

Surveys become more valuable

"...the abundance of [large digital] data sources increases—not decreases—the value of surveys." (Salganik, 2015, p. 87)

Total Survey Error, Bias, Variance

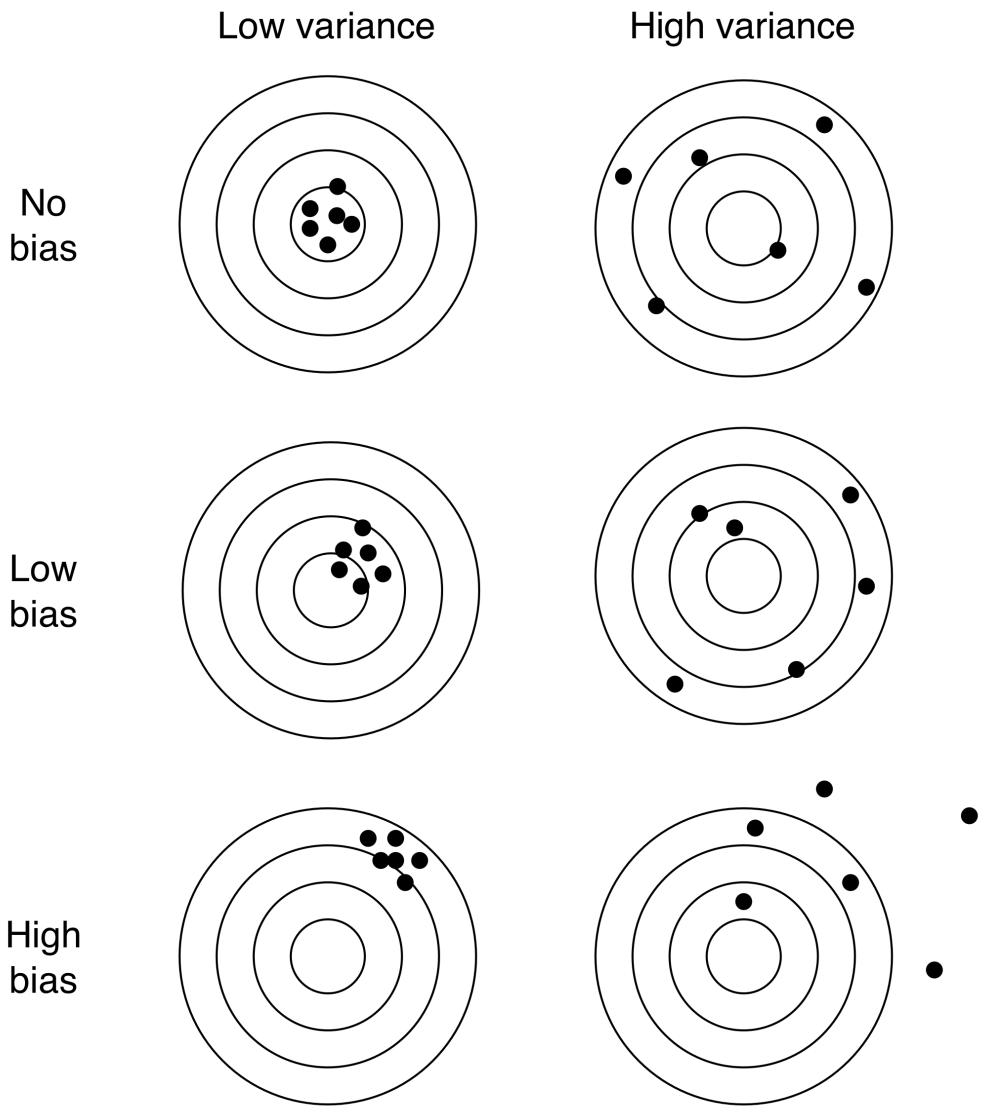
$$\text{Total Survey Error} = \text{Representation Errors} + \text{Measurement Errors}$$

$$\text{Total Survey Error} = \text{Bias} + \text{Variance}$$

Two objectives

- You don't want to minimize one and allow lots of the other
- Best inference balances these two objectives
- Related topic: overfitting

Total Survey Error, Bias, Variance

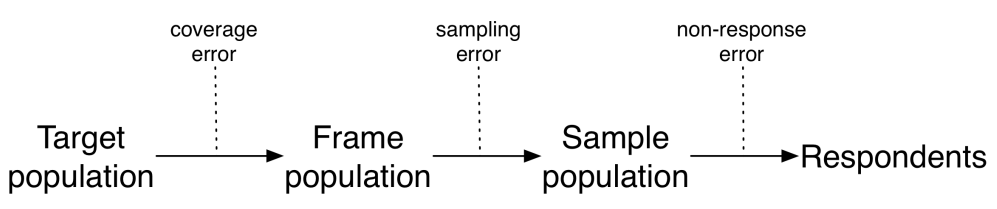


Total Survey Error, Bias, Variance

Words of sampling wisdom

"Perfect sampling with bad survey questions will produce bad estimates, as will bad sampling with perfect survey questions." (Salganic, 2015, pp.90-91)

Representativeness



- target population: population of interest
- frame population: list of people that can be used for sampling
- sample population: people from frame population who researcher attempts to interview
- Respondents: members of the sample population who respond to the survey

Representativeness

Two lessons from unrepresentative fiascoes

1. Having a large number of respondents will often decrease the variance of estimates, but it does not necessarily decrease the bias.
2. Researchers need to account for how their sample is collected in making estimates—how might the population be skewed.

Measurement

How you ask matters: question form effects

Two priests, a Dominican and a Jesuit, are discussing whether it is a sin to smoke and pray at the same time. After failing to reach a conclusion, each goes off to consult his respective superior. The Dominican says, “What did your superior say?”

The Jesuit responds, “He said it was alright.”

“That’s funny.” the Dominican replies, “My supervisor said it was a sin.”

The Jesuit said, “What did you ask him?” The Dominican replies, “I asked him if it was alright to smoke while praying.” “Oh” said the Jesuit, “I asked if it was OK to pray while smoking.”

Measurement

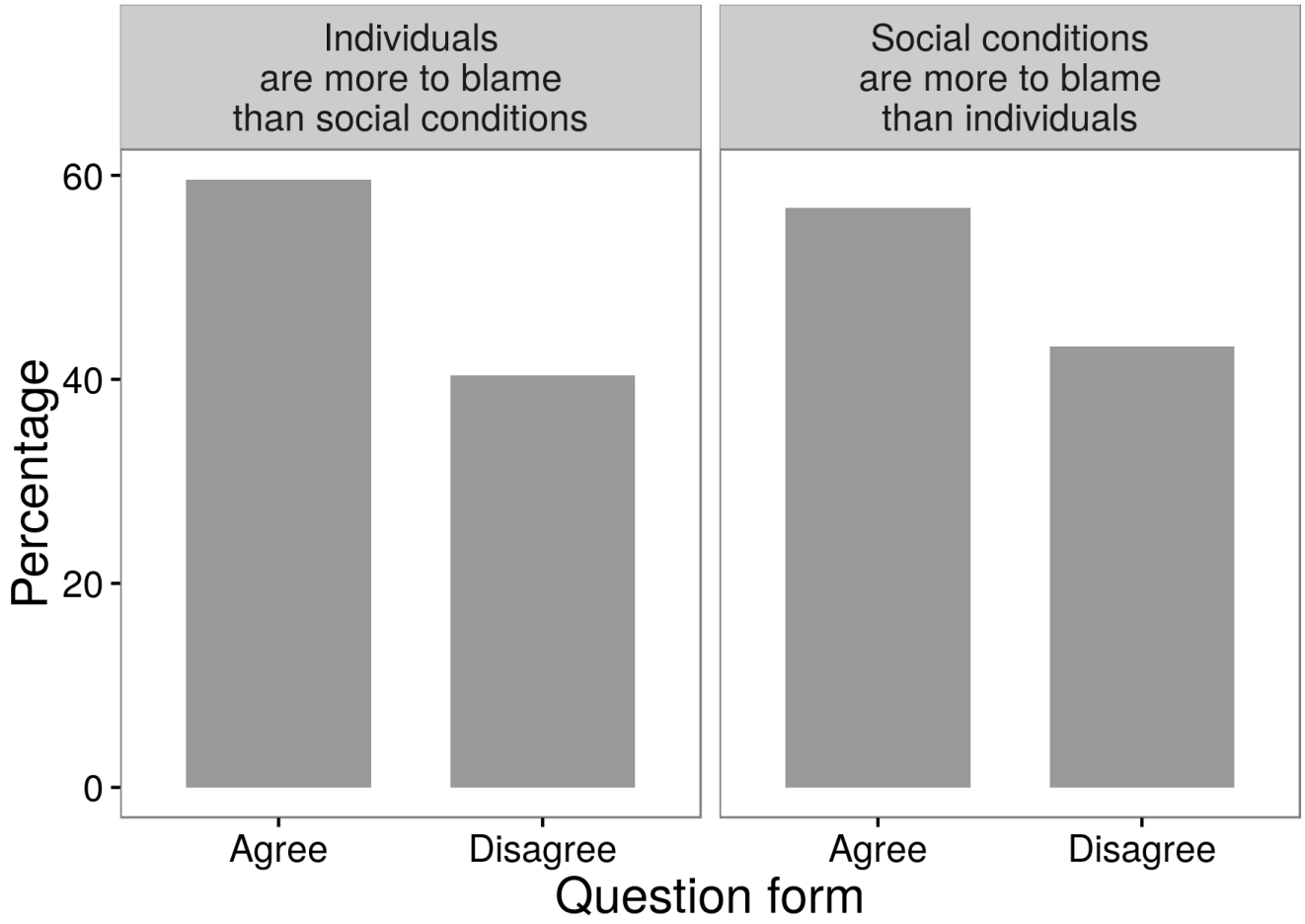


Fig. 3.3 from Salganik (2018), adapted from Schumann and Presser (1996, table 8.1).

Measurement

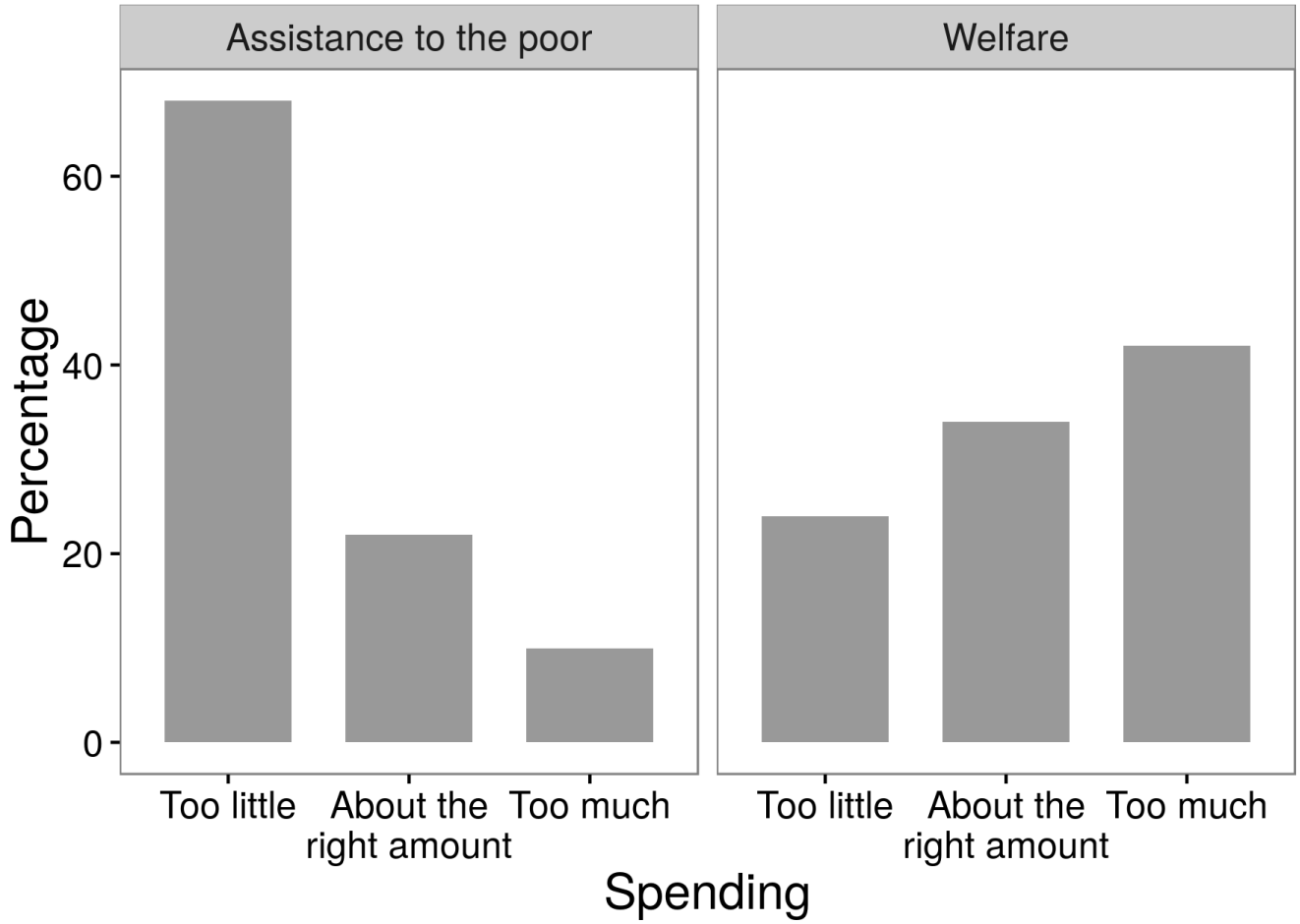


Fig. 3.4 from Salganik (2018), adapted from Huber and Paris (2013, table A1).

Measurement

How might you ask a question to get someone's age?

- How old are you?
- What is your age?
- What is your age? (asked as the first question)
- What is your age? (asked as the last question)
- What year were you born?
- Were you born before or after 1975?

Measurement suggestions

- Bradburn, Sudman, and Wansink (2004), *Asking Questions: The Definitive Guide to Questionnaire Design—For Market Research, Political Polls, and Social and Health Questionnaires*, rev. ed., San Francisco: Jossey-Bass.
- Copy—word for word—questions from high-quality surveys
- Survey experiment some of your questions
  - Half respondents get one wording, half get other wording
- Pilot test your questions: pre-testing
  - Like a survey experiment with a small initial part of your sample

Design survey from scratch

- **Research Question:** How does cell phone use affect high school test scores?
  - How would we design a survey?
  - What questions?
  - What post stratification issues?
  - What biases?

Canann and Evans (2015), Payday lenders

Research Question

Do payday and title lenders price discriminate based on socioeconomic and demographic characteristics?

- Survey payday and title lenders in Utah
- Add zip code and MSA level data

Assignment 4

- Look through Assignment 4.

Probability and non-probability sampling methods

- **Non-probability sampling:** not all members of target population have known, nonzero probability of being sampled.
- **Probability sampling:** all members of target population have known, nonzero probability of being sampled.
  - Can easily re-weight sample and respondent populations to be representative
  - provable properties about inference accuracy
  - Currently dominant approach in social scientific research

Problems with probability sampling

- Theoretical assumptions on which proofs are based are often violated
  - Coverage errors
  - Nonresponse rates (steadily rising)

Non-response rates

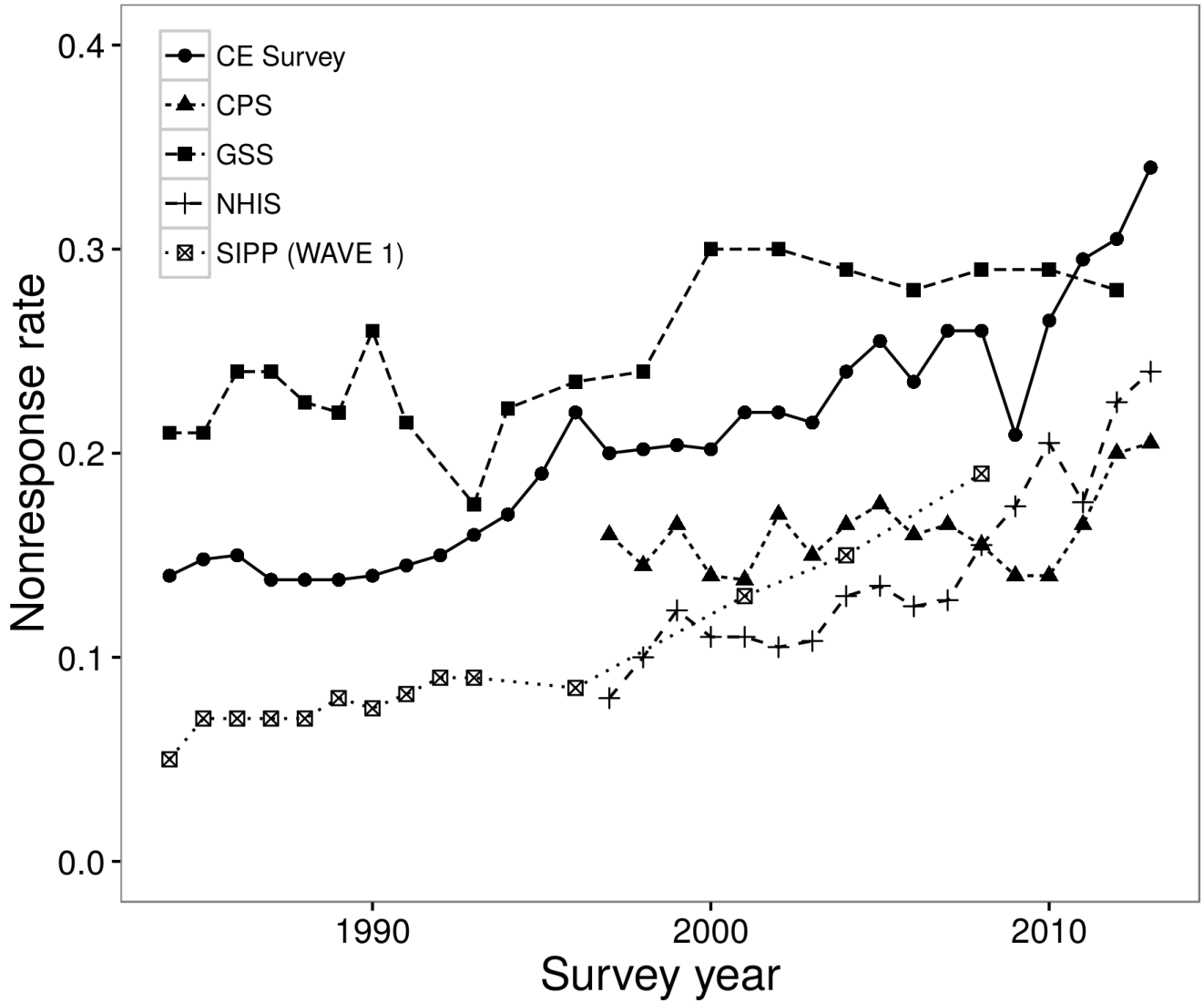


Fig. 3.5 from Salganik (2018), adapted from Meyer, Mok, and Sullivan (2015, figure 1).

Probability Sampling Failures

Literary Digest straw poll U.S. Presidential elections

- Correctly predicted 1920, 1924, 1928, 1932
- Mail ballots to 10 million people

"THE DIGEST's smooth-running machine moves with the swift precision of thirty years' experience to reduce guesswork to hard facts.... This week 500 pens scratched out more than a quarter of a million addresses a day. Every day, in a great room high above motor-ribboned Fourth Avenue, in New York, 400 workers deftly slide a million pieces of printed matter—enough to pave forty city blocks—into the addressed envelops [sic]. Every hour, in THE DIGEST'S own Post Office Substation, three chattering postage metering machines sealed and stamped the white oblongs; skilled postal employees flipped them into bulging mailsacks; fleet DIGEST trucks sped them to express mail-trains.... Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be triple-checked, verified, five-times cross-classified and totaled. When the last figure has been totted and checked, if past experience is a criterion, the country will know to within a fraction of 1 percent the actual popular vote of forty million [voters]." (August 22, 1936)

Probability Sampling Failures

- 1936 *Literary Digest* straw poll prediction
  - 2.4 million mailer ballots returned (24% response rate)
  - Alf Landon should defeat incumbent Franklin Roosevelt
- Roosevelt wins (60.8% popular, 523/531 electoral college)

Problems

- Sampled telephone directories and automobile registrations
- Predominantly wealthier
- Challenger (Landon) supporters were more likely to respond

Probability Sampling Failures

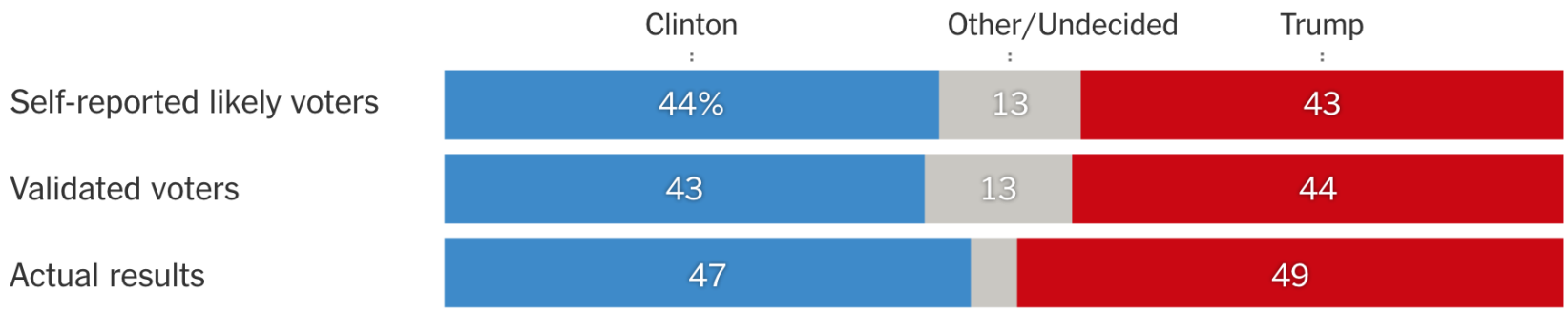
Fig. 3.6 from Salganik (2018), “Dewey Defeats Truman” headline from 1948 was based on nonrepresentative survey problem.

Probability Sampling Failures

- 2016 U.S. Presidential Election: Clinton vs. Trump
- See Nate Cohn *NYT, 5/31/17* article, "A 2016 Review: Why Key State Polls Were Wrong About Trump"
  - "Undecided voters broke for Mr. Trump in the final days of the race"
  - "Turnout among Mr. Trump's supporters was somewhat higher than expected"
  - State polls "understated Mr. Trump's support in the decisive Rust Belt region, in part because those surveys did not adjust for the educational composition of the electorate"

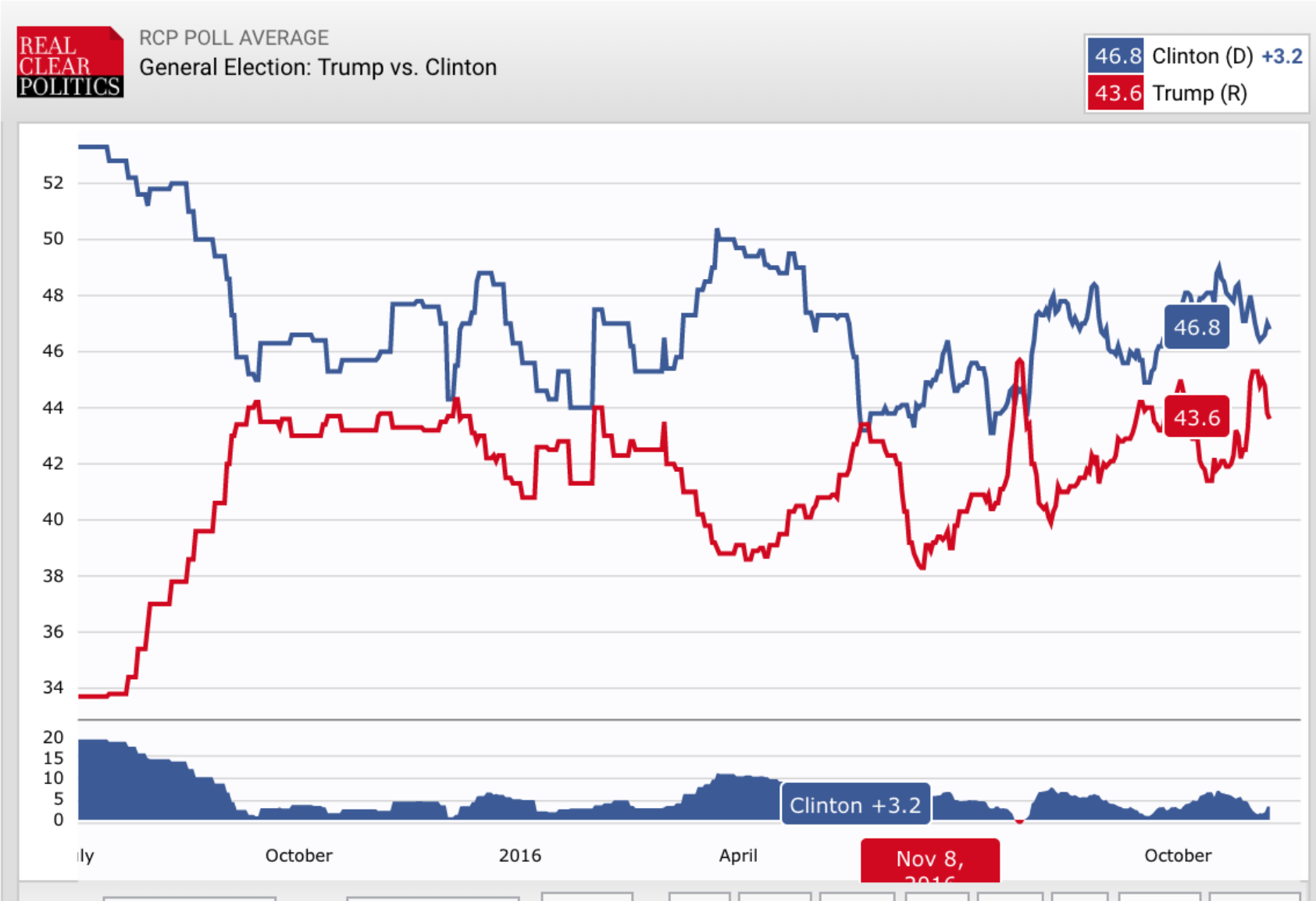
Trump vs. Clinton 2016

New York Times Upshot/Siena poll results in final Fla., Pa. and N.C. polls among self-reported likely voters and those who actually voted.



Source: Upshot/Siena polls and voter history data from the Florida Secretary of State, Pennsylvania Secretary of State and North Carolina Board of Elections. Self-reported likely voters said they had already voted, were almost certain to vote or very likely to vote.

Trump vs. Clinton 2016



Non-probability sampling methods

Non-probability sampling approach

- Probability samples are becoming more difficult to obtain
- **Post stratification**: Do non-probability sampling, then re-weight according to target population

- See Wang, Rothschild, Goel, and Gelman (2015), Assignment 4
- Use Xbox users data and post stratification to successfully predict 2012 U.S. election: Romney vs. Obama

Keys and problems to post stratification

Keys to post stratification

- Have the right groups
- Represent important heterogeneity in target population
- homogeneous response propensities within groups

Problems

- Too many groups can make too little data
- Small bin size
- Might only have one female, Asian, manufacturing workers

Digital advances in survey approaches

1. Recruiting respondents
2. How to ask questions
3. Combining with large digital data

Eras of survey modes

- face-to-face
- mail
- telephone
- smart phones and computers

Computer administered surveys

Benefits

- Reduce costs
- Reduce social desirability bias (respondents underreporting stigmatized behavior)
- Eliminate interviewer effects (recording responses influenced by interviewer)
- Increase respondent flexibility, timing

Drawbacks

- Cannot clarify confusing questions (AI?)
- Lose rapport with respondent
- Interviewer can help maintain respondent engagement (EMA, gamification)

Ecological Momentary Assessments

"Ecological momentary assessment (EMA) involves taking traditional surveys, chopping them up into pieces, and sprinkling them into the lives of participants." (Salganik, 2018, p. 109)

- collection of data in real-world environments
- assessments focus on respondents' current or very recent states
- assessment may be event-based, time-based, or randomly prompted
- completion of multiple assessments over time

EMA: Sugie (2014, 2016)

Sugie, Naomi F., "Finding Work: A Smartphone Study of Job Searching, Social Contacts, and Wellbeing After Prison," PhD Thesis, Princeton University (2014).

- Complete list of individuals leaving prison in Newark, NJ
- Gave them smartphone
- Asked questions at regular and random times
- Combined with location and accelerometer data

Ethical concerns

- Got IRB approval
- Went beyond: informed consent from each participant
- Went beyond: ability to turn off and on tracking
- Went beyond: Certificate of Confidentiality from government

EMA: Sugie (2014, 2016)

Sugie (2014) findings

- Participants work experience mostly informal
- Early exit pattern
- Persistent search pattern
- Recurring work pattern
- Low response pattern
- Early exit/stop searching group (least likely to be successful)

EMA: York Cornwell and Cagney (2017)

York Cornwell, Erin and Kathleen A. Cagney, "Aging in Activity Space: Results From Smartphone-Based GPS-Tracking of Urban Seniors," *The Journals of Gerontology: Series B* 72:5, (Sep. 2017) pp. 864-875.

- Studies the importance of neighborhood in older adults
- Gave smartphone to 60 older adults in New York City
- Administered 17 EMA surveys over 4 days
- Logged location every 5 minutes

EMA: York Cornwell and Cagney (2017)

	Respondent		
	A	B	C
Observations outside of residential tract	36.3%	11.4%	65.0%
Area of standard deviation ellipse	1.00 mi^2	.74 mi^2	8.60 mi^2
Total tracts visited	13	21	86
Tracts visited for 10+ minutes	9	4	13
Poverty rate in residential tract	31.5%	33.5%	34.5%
Average poverty rate in non-residential tracts	30.8%	13.6%	23.8%

Wiki Surveys: Closed vs. Open responses

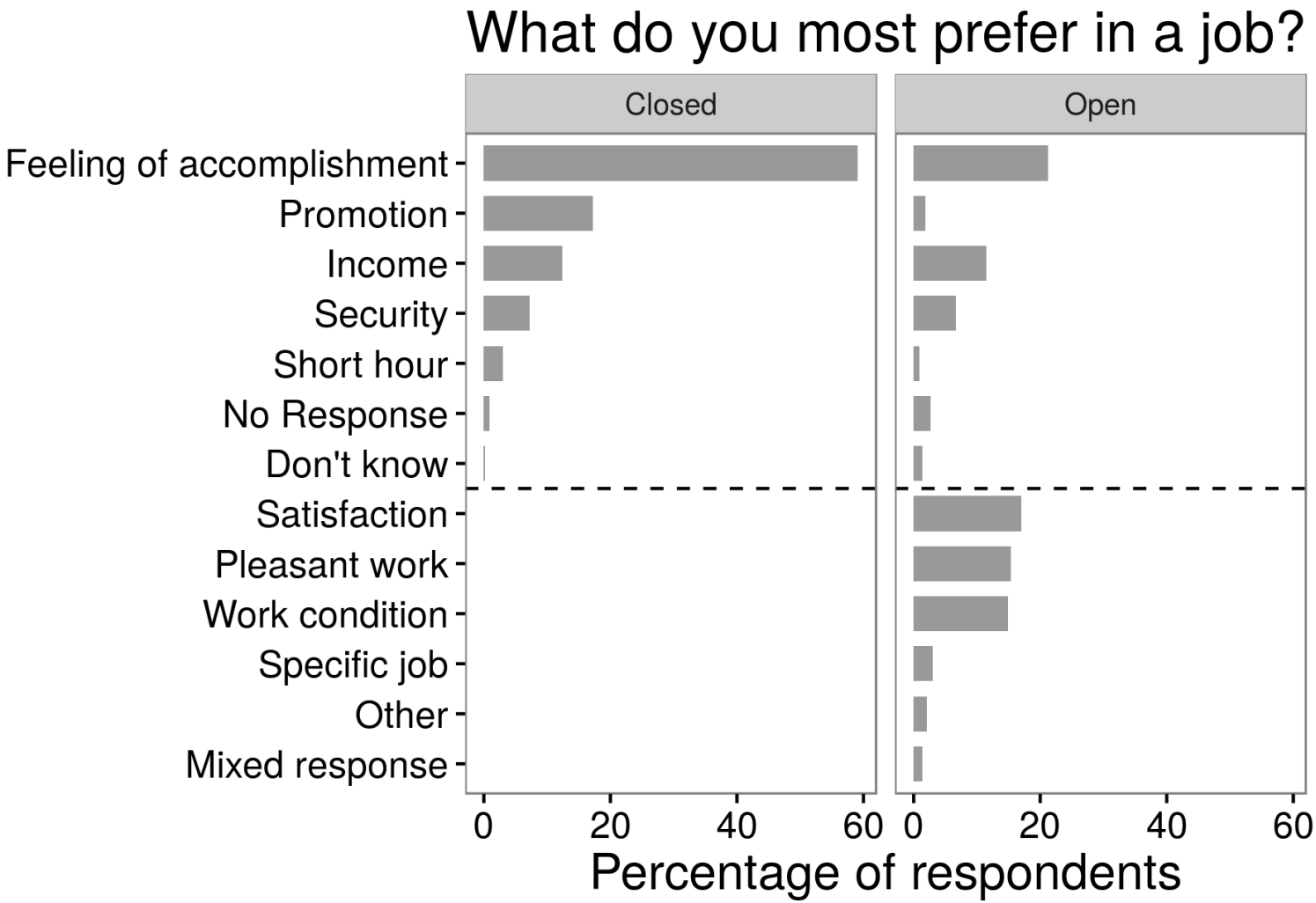


Fig. 3.9 from Salganik (2018), adapted from Shuman and Presser (1979, table 1).

Wiki Surveys

Open vs. closed questions

- Closed questions dominate social science surveys
- Closed are easier to use
- Open questions are not uniform format
- Open questions are expensive to analyze
- Analyzing open questions is error-prone

But...

Open questions usually give you exactly the information you want

Wiki Surveys

Combine closed and open questions and include methods that allow analysis of open questions. www.allourideas.org

Gamification

With computer interviewer

- Respondents might lose interest and leave participation
- **Gamification**: make process of responding more enjoyable

Goel, Mason, Watts (2010)

- Facebook app to test: perception versus reality of similarity to friends
- Game with/against friends. How well do you know?

Gamification

Goel, Mason, Watts (2010) findings

- Friends more likely to give same answer than strangers
- Even close friends disagree on 30% of questions
- Respondents overestimated agreement with friends
  - Implication: most diversity of opinions among friends is unnoticed
- Respondents equally likely to be aware of disagreements with friends on serious matters and lighthearted matters.

- How could we gamify sleep number data?
- How could we gamify our high school cell phone use and intelligence survey?

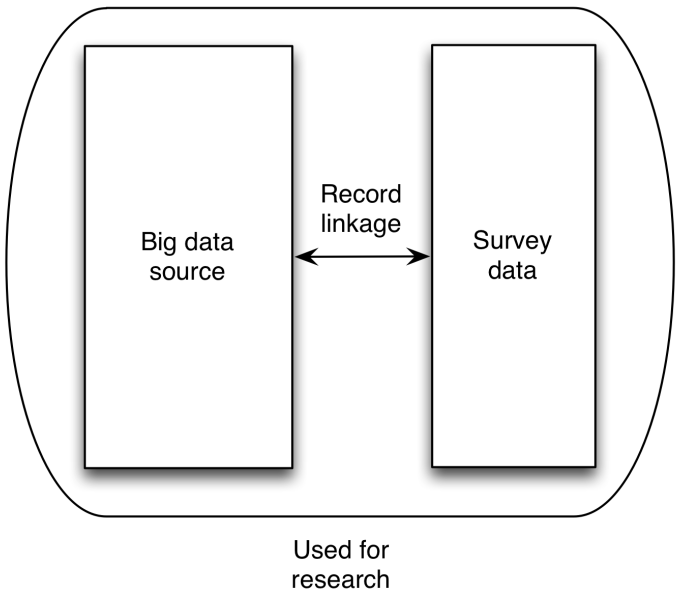
Surveys linked to large digital data

Two main ways to link surveys to large data

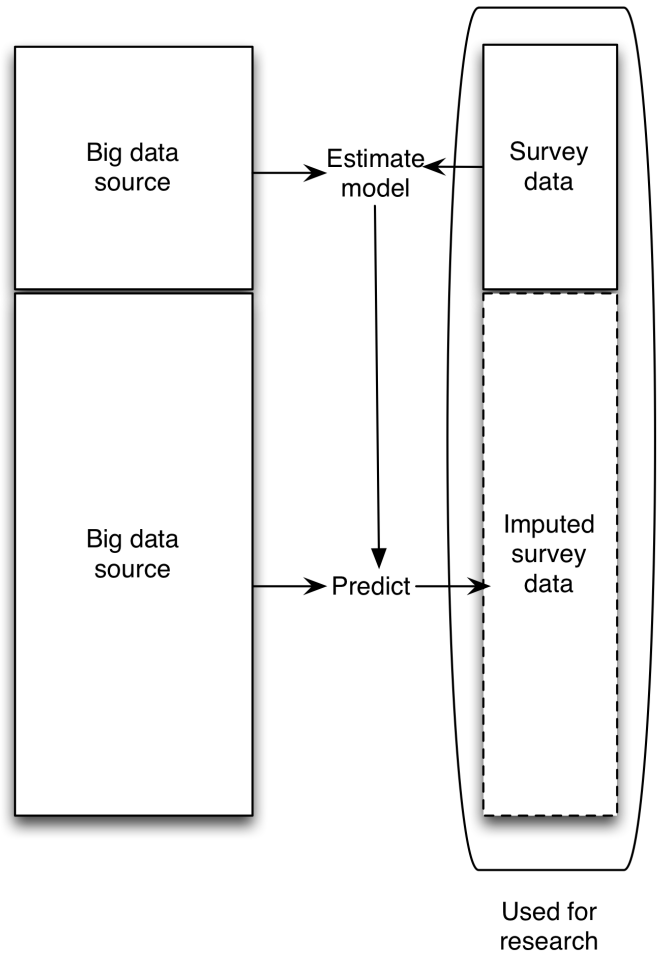
1. Enriched asking
   - Link surveys to large data sources
2. Amplified asking
   - Link survey to portion of large data
   - Estimate/fit model of relationship between large data and survey
   - Use model to impute survey data on rest of large data set

Surveys linked to large digital data

Enriched asking



Amplified asking



Enriched Asking

- Large digital data has many observations
- Lacks other measurements
- Link survey data with large observational data

Difficulties

- Record linkage difficult if no unique identifiers
- Matching based on observable characteristics

Enriched Asking: Ansolabehere and Hersh (2012)

Ansolabehere, Stephen and Eitan Hersh (2012), "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate," *Political Analysis*, 20:3, pp. 437-459.

- Large digital voting records from Catalist LCC
- Merged survey demographic and attitudes with voting record data
- Found that over-reporting of voting is rampant (20% reporting "voted" didn't vote)
- Over reporting more common among high-income, high-education, partisans, engaged in public affairs
- Actual differences between voters and nonvoters smaller than appear

Amplified Asking

- Combine small amount of data with large digital data source
- Estimate relationship (model) of observational features (variables) to survey features (variables)
- Use model to predict (impute) survey data on rest of observational data

Difficulties

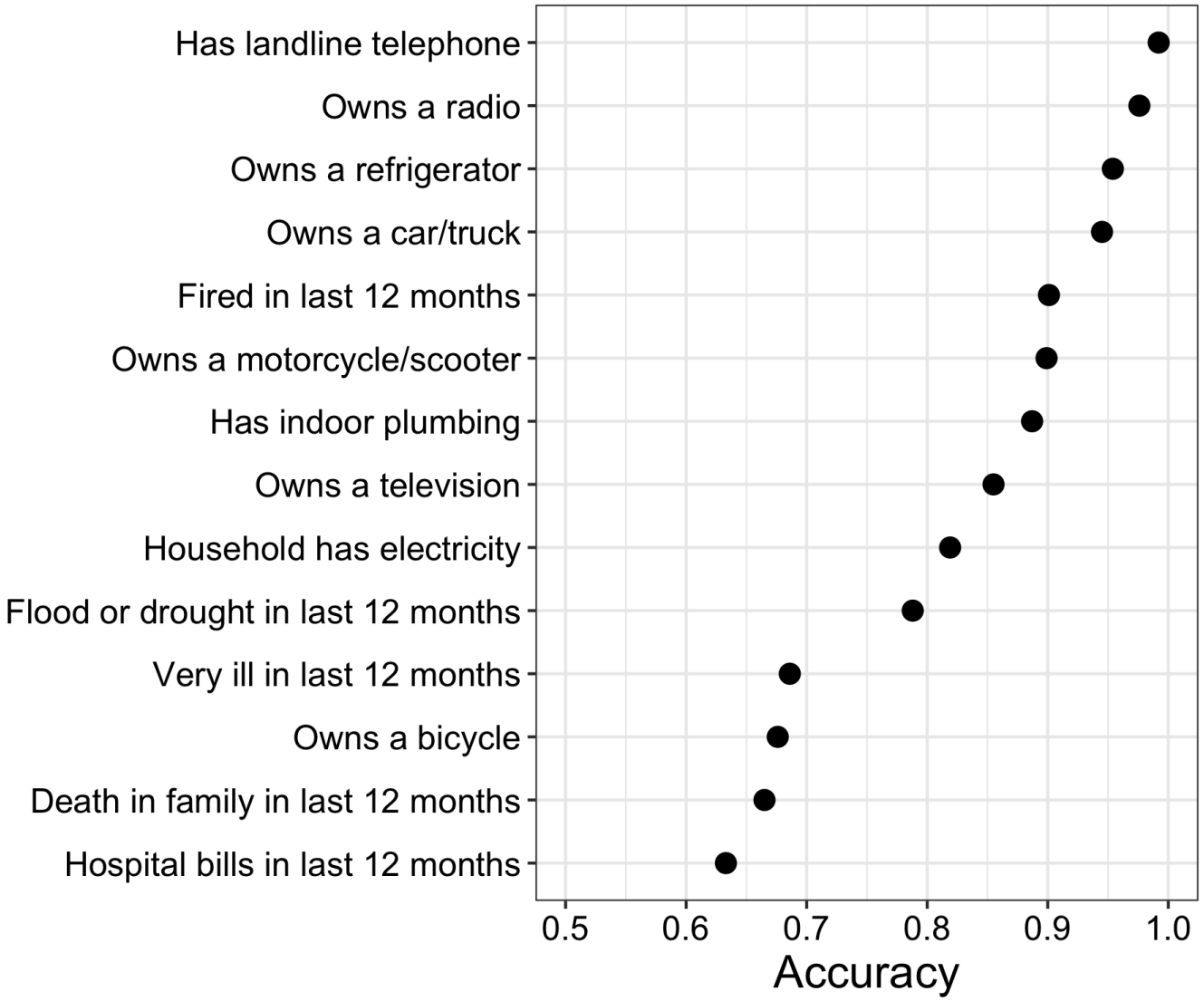
- How good is predictive imputation?
- How representative is final sample? Is broader inference valid?

Amplified Asking: Blumenstock, et al (2012)

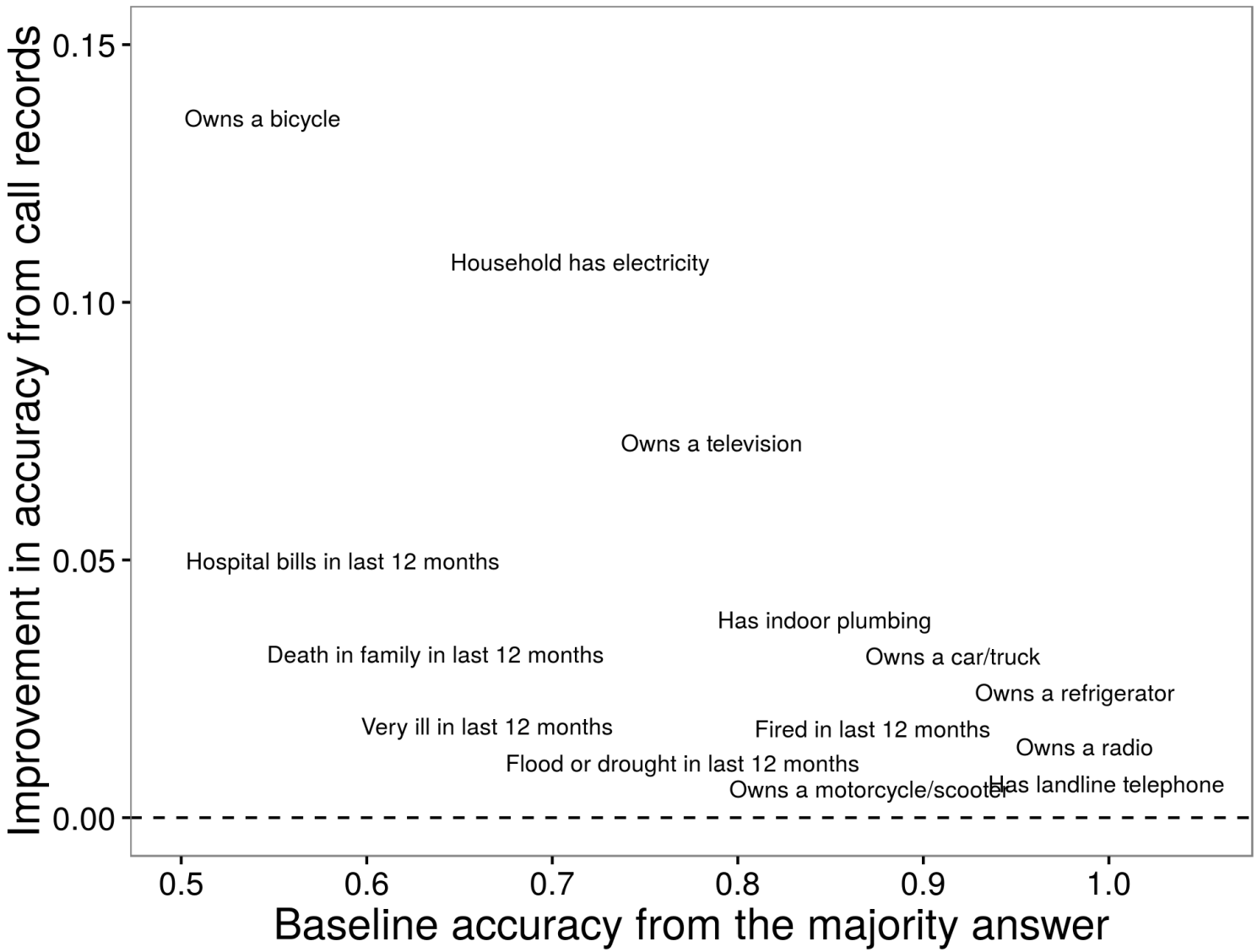
Blumenstock, Joshua (2014), "Calling for Better Measuremenet: Estimating an Individual's Wealth and Well-Being from Mobile Phone Transaction Records," Presented at KDD—Data Science for Social Good 2014, New York.

- 1.5 million cell customers in Rwanda from 2005 to 2009
- Call and text message data: time, duration, location
- Survey small sample: Do you own a radio?, bicycle?, refrigerator?, etc.
- Estimate model relationship between observational data and survey data
- Impute survey data on rest of observational data
- Results: predicted wealth geography accurately and cheaply

Amplified asking: Blumenstock (2014)



Amplified asking: Blumenstock (2014)



Surveys: final flourish

Evans and Berman discussion: “Determinants of tax policy preferences and their associated effects?”

- Have model with different tax policies and effects: TaxBrain and Tax-Calculator
- Survey individuals to have them choose a tax policy and observe its effects
- Maybe can change policy through experimentation
- Other survey responses: age, gender, religion, party affiliation, income, geography,

Ways to get data

- What are some good ways to get these data?
- How can I make sure it is representative?