Survey Research and Design

Bird's-Eye View of Survey Research

William Marble

September 7, 2022

Housekeeping

- Paper presentation sign-up — look for Canvas announcement this week
- Slides on Canvas
- R setup and installation — finish by next class

Roadmap

1. What is a survey?
2. A bottom-up view of survey research
3. A top-down view of survey research
4. Discussion of survey error in Lee and Zhang (2017)

What is a Survey?

What is a Survey?

A survey is a **systematic** method for gathering information from (a **sample** of) entities for the purposes of constructing **quantitative descriptors** of the **attributes** of the larger **population** of which the entities are members. (Groves, 2)

What is a Survey?

A survey is a **systematic** method for gathering information from (a **sample** of) entities for the purposes of constructing **quantitative descriptors** of the **attributes** of the larger **population** of which the entities are members. (Groves, 2)

Systematic: standardized, following a set procedure

What is a Survey?

A survey is a **systematic** method for gathering information from (a **sample** of) entities for the purposes of constructing **quantitative descriptors** of the **attributes** of the larger **population** of which the entities are members. (Groves, 2)

Systematic: standardized, following a set procedure

Sample: a subset that doesn't contain every eligible unit

What is a Survey?

A survey is a **systematic** method for gathering information from (a **sample** of) entities for the purposes of constructing **quantitative descriptors** of the **attributes** of the larger **population** of which the entities are members. (Groves, 2)

Systematic: standardized, following a set procedure

Sample: a subset that doesn't contain every eligible unit

Quantitative descriptor: a summary of information gathered in terms of a statistic. e.g. means, standard deviations, correlations between variables

What is a Survey?

A survey is a **systematic** method for gathering information from (a **sample** of) entities for the purposes of constructing **quantitative descriptors** of the **attributes** of the larger **population** of which the entities are members. (Groves, 2)

Systematic: standardized, following a set procedure

Sample: a subset that doesn't contain every eligible unit

Quantitative descriptor: a summary of information gathered in terms of a statistic. e.g. means, standard deviations, correlations between variables

Attributes: the concepts that we are ultimately interested in. e.g. presidential approval, income inequality

What is a Survey?

A survey is a **systematic** method for gathering information from (a **sample** of) entities for the purposes of constructing **quantitative descriptors** of the **attributes** of the larger **population** of which the entities are members. (Groves, 2)

Systematic: standardized, following a set procedure

Sample: a subset that doesn't contain every eligible unit

Quantitative descriptor: a summary of information gathered in terms of a statistic. e.g. means, standard deviations, correlations between variables

Attributes: the concepts that we are ultimately interested in. e.g. presidential approval, income inequality

Population: the universe of units that we are ultimately interested in. e.g. all people in the U.S., all businesses with at least 10 employees, registered voters in Pennsylvania

What is a Survey?

A survey is a **systematic** method for gathering information from (a **sample** of) entities for the purposes of constructing **quantitative descriptors** of the **attributes** of the larger **population** of which the entities are members. (Groves, 2)

Systematic: standardized, following a set procedure

Sample: a subset that doesn't contain every eligible unit

Quantitative descriptor: a summary of information gathered in terms of a statistic. e.g. means, standard deviations, correlations between variables

Attributes: the concepts that we are ultimately interested in. e.g. presidential approval, income inequality

Population: the universe of units that we are ultimately interested in. e.g. all people in the U.S., all businesses with at least 10 employees, registered voters in Pennsylvania

Our Goals in this Course

- Ultimately, surveys are useful because there is a concept in the world that we would like to learn
- The field of survey methodology is aimed at developing tools to obtain an accurate **estimate**
- **Error** is unavoidable — we'll develop tools to categorize, understand, minimize, and quantify it
- By the end of the semester we'll have many tools

Survey Methodology is Interdisciplinary

We'll be wearing a lot of hats:

- statistician
- data scientist
- writer
- subject-matter expert
- cognitive scientist

Survey Methodology is Evolving

Survey methodology is changing quickly right now

- Classic, well-established methods are becoming more expensive and less reliable
- New methods are actively being developed in this field

Survey Methodology is Evolving

Survey methodology is changing quickly right now

- Classic, well-established methods are becoming more expensive and less reliable
- New methods are actively being developed in this field

Our approach:

- First cover classic methods: baseline, "idealized" process of survey research
- Then cover departures from the ideal and (potential) solutions

Survey Design from the Bottom Up

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)
6. Draw a **sample** from the sampling frame

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)
6. Draw a **sample** from the sampling frame
7. Analyze the data (**analysis**)

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)
6. Draw a **sample** from the sampling frame
7. Analyze the data (**analysis**)
   - Clean and recode data

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)
6. Draw a **sample** from the sampling frame
7. Analyze the data (**analysis**)
   - Clean and recode data
   - Weight to target population

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)
6. Draw a **sample** from the sampling frame
7. Analyze the data (**analysis**)
   - Clean and recode data
   - Weight to target population
   - Compute averages and crosstabs

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)
6. Draw a **sample** from the sampling frame
7. Analyze the data (**analysis**)
   - Clean and recode data
   - Weight to target population
   - Compute averages and crosstabs
   - Visualize results

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)
6. Draw a **sample** from the sampling frame
7. Analyze the data (**analysis**)
   - Clean and recode data
   - Weight to target population
   - Compute averages and crosstabs
   - Visualize results
   - ...

Survey Lifecycle in Practice

1. Identify topics to investigate (**constructs**)
2. Write survey questions to measure the construct (**measurement**)
3. Define response type and options (**response**)
4. Identify universe of units to investigate (**population**)
5. Identify people who may be included in the survey (**sampling frame**)
6. Draw a **sample** from the sampling frame
7. Analyze the data (**analysis**)
   - Clean and recode data
   - Weight to target population
   - Compute averages and crosstabs
   - Visualize results
   - ...
8. Write up results (**reporting**)

Survey Design in Practice

- There are lots of design decision to make:
  - how to word questions? what response options to give?
  - which sample frame?
  - how to recruit sample?
  - whether and how to compute weights?
  - etc.
- We need a framework to help guide these decisions

Survey Design from the Top Down

Survey Design from the Top Down

- Instead of thinking about the bottom-up *process*, a top-down scheme emphasizes **total survey error**
- We have our idealized, *unobservable* target of inference and our observable operationalization
- Each step of the survey process may introduce error: differences between the target of inference and our measurement

A Running Example

Suppose we want to measure the political ideology of registered voters in Philadelphia.

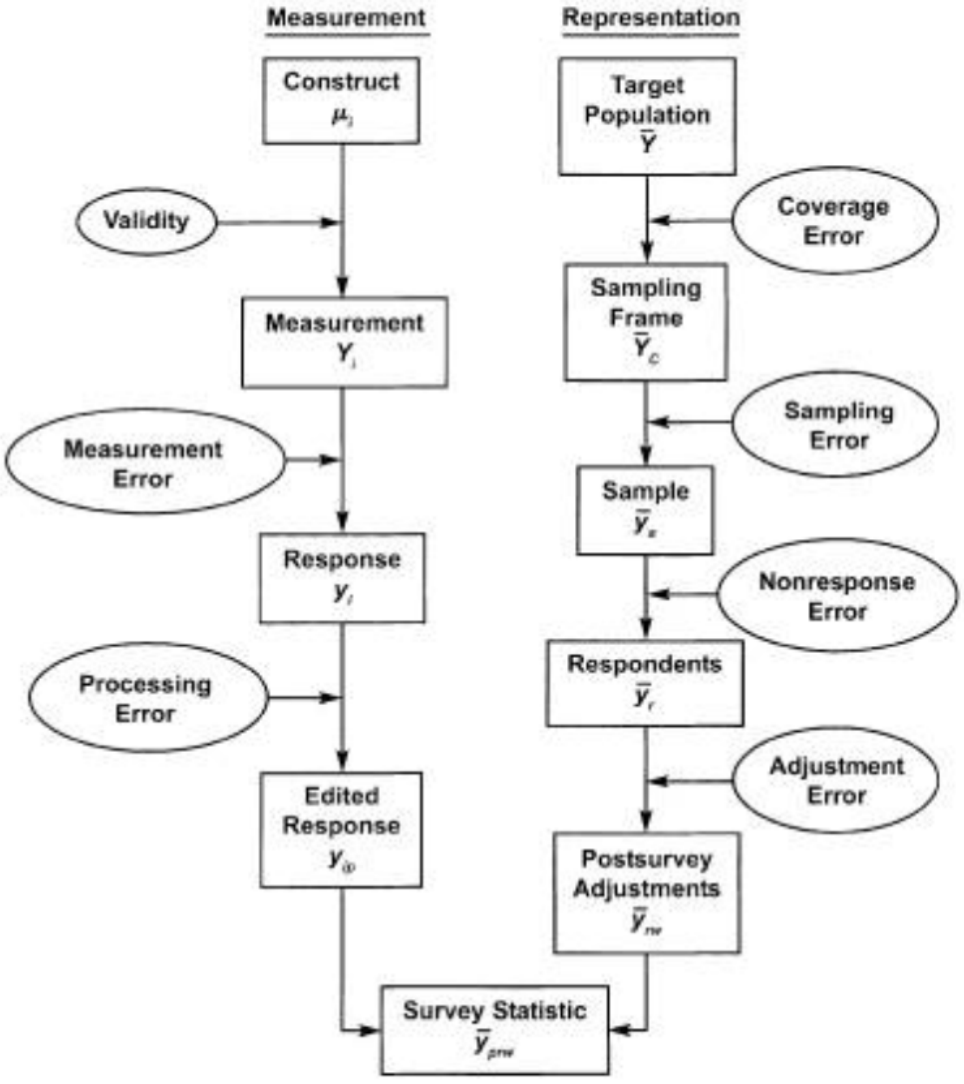


Figure 2.5 Survey life cycle from a quality perspective.

Defining the Construct: What Are We Trying to Measure?

Defining the Construct: What Are We Trying to Measure?

Presidential job approval

Defining the Construct: What Are We Trying to Measure?

Presidential job approval

- what does it mean to "approve"?
- what components of the "job" do we care about assessing?

Defining the Construct: What Are We Trying to Measure?

Presidential job approval

- what does it mean to "approve"?
- what components of the "job" do we care about assessing?

Unemployment

Defining the Construct: What Are We Trying to Measure?

Presidential job approval

- what does it mean to "approve"?
- what components of the "job" do we care about assessing?

Unemployment

- over what time period?
- what counts as employment?

Defining the Construct: What Are We Trying to Measure?

Presidential job approval

- what does it mean to "approve"?
- what components of the "job" do we care about assessing?

Unemployment

- over what time period?
- what counts as employment?

Partisanship

Defining the Construct: What Are We Trying to Measure?

Presidential job approval

- what does it mean to "approve"?
- what components of the "job" do we care about assessing?

Unemployment

- over what time period?
- what counts as employment?

Partisanship

- registration?
- as social identity?
- typical vote choice?

Ideology in Philly: Construct

What is ideology?

Ideology in Philly: Construct

What is ideology?

- are liberal-conservative the ideologies we care about?
- how do libertarianism or Marxism fit?
- does ideology have policy content or is policy downstream of ideology?

Ideology in Philly: Construct

What is ideology?

- are liberal-conservative the ideologies we care about?
- how do libertarianism or Marxism fit?
- does ideology have policy content or is policy downstream of ideology?

We'll define ideology as a worldview that organizes one's political beliefs.

From Construct to Measure

- We've developed our construct μ_i . . . Now we need to measure it

From Construct to Measure

- We've developed our construct μ_i . . . Now we need to measure it
- Problem: Often we cannot directly measure μ_i even in theory

From Construct to Measure

- We've developed our construct μ_i . . . Now we need to measure it
- Problem: Often we cannot directly measure μ_i even in theory
- Solution: We come up with an observable measure Y_i

From Construct to Measure

- We've developed our construct μ_i . . . Now we need to measure it
- Problem: Often we cannot directly measure μ_i even in theory
- Solution: We come up with an observable measure Y_i

Validity

- The extent to which the measure Y_i is the same as the construct itself is called **validity**
- In our notation, we can think of validity as $\varepsilon_i = \mu_i - Y_i$
- When ε_i is low, we have high validity, and vice versa

Ideology in Philly: Validity

If ideology μ_i is an organizing worldview, the measure Y_i could be whether the respondent thinks of themselves as liberal or conservative:

> *In general, how would you describe your own political viewpoint? Very liberal, Liberal, Moderate, Conservative, or Very Conservative?*

Ideology in Philly: Validity

If ideology μ_i is an organizing worldview, the measure Y_i could be whether the respondent thinks of themselves as liberal or conservative:

> *In general, how would you describe your own political viewpoint? Very liberal, Liberal, Moderate, Conservative, or Very Conservative?*

Under what assumptions is this a valid measure ideology? Why might it not be?

From Measure to Measurement

- After we've decided on an observable **measure**, we have to collect data
- For surveys, we ask people a question and record their response
- Groves denotes an observed survey response y_{it}
- Note t subscript because we imagine the same person repeatedly answering the question (even if in reality they only answer once)
- Differences between Y_i and y_{it} is **measurement error**

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

- Respondents make mistakes

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

- Respondents make mistakes
- Respondents might lie

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

- Respondents make mistakes
- Respondents might lie
- We might give coarse response options

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

- Respondents make mistakes
- Respondents might lie
- We might give coarse response options
- Respondent might be uncertain about how to map ideas in their head onto the options we give them

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

- Respondents make mistakes
- Respondents might lie
- We might give coarse response options
- Respondent might be uncertain about how to map ideas in their head onto the options we give them

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

- Respondents make mistakes
- Respondents might lie
- We might give coarse response options
- Respondent might be uncertain about how to map ideas in their head onto the options we give them

If on average y_{it} is the same as Y_i we say the measure is **unbiased**: $E(y_{it}) = Y_i$

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

- Respondents make mistakes
- Respondents might lie
- We might give coarse response options
- Respondent might be uncertain about how to map ideas in their head onto the options we give them

If on average y_{it} is the same as Y_i we say the measure is **unbiased**: $E(y_{it}) = Y_i$

- what do we mean by "average"?

From Measure to Measurement

We can imagine the same person giving slightly different answers to the same question. (why?)

- Respondents make mistakes
- Respondents might lie
- We might give coarse response options
- Respondent might be uncertain about how to map ideas in their head onto the options we give them

If on average y_{it} is the same as Y_i we say the measure is **unbiased**: $E(y_{it}) = Y_i$

- what do we mean by "average"?
- thought experiment: imagine we ask the question infinitely many times (to different people)

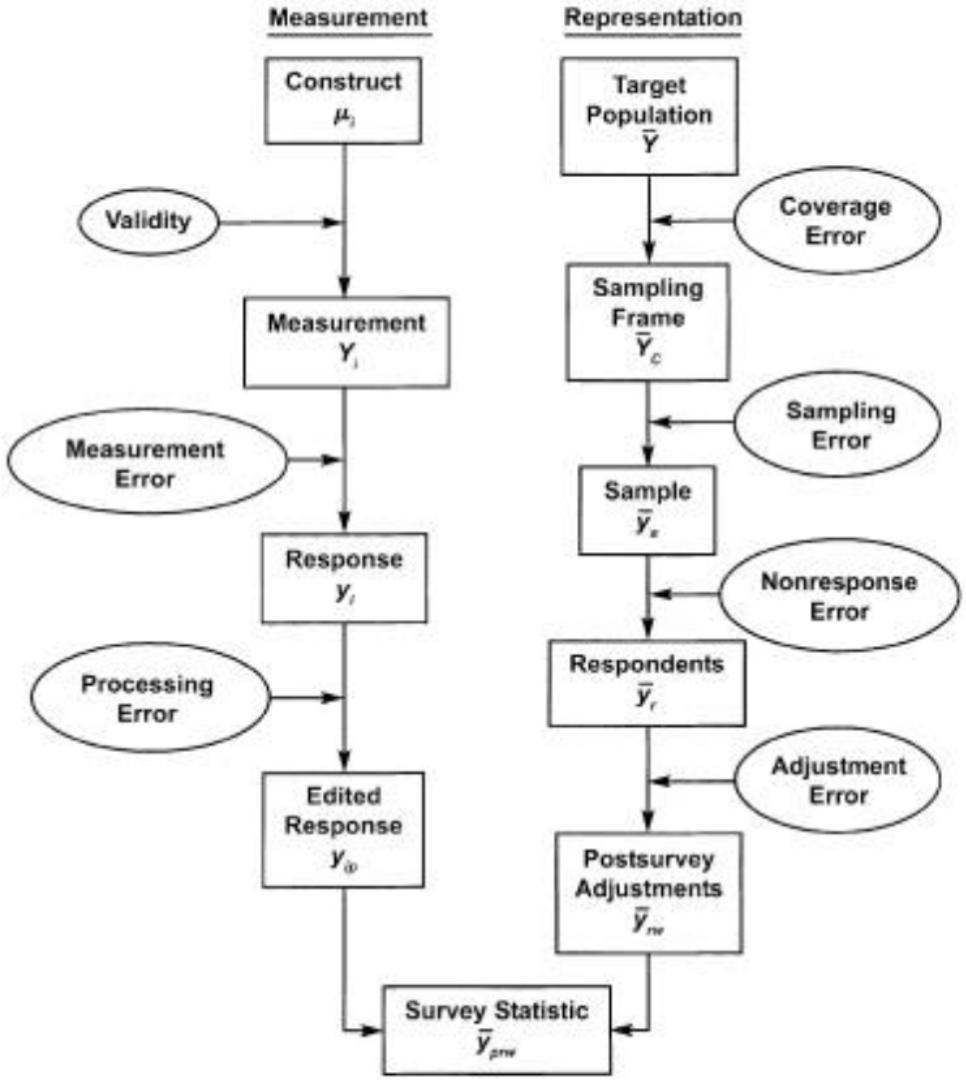


Figure 2.5 Survey life cycle from a quality perspective.

Defining the Population: Who Do We Care About?

- We've decided on a concept and measurement strategy
- Now we need to decide who we're interested in
- This is the **target population**
- Often the target population is obvious based on research question

Defining the Population: Who Do We Care About?

- We've decided on a concept and measurement strategy
- Now we need to decide who we're interested in
- This is the **target population**
- Often the target population is obvious based on research question
- Typical default choices for public opinion research:
  - Adult population
  - Adult citizen population
  - Registered voters

Defining the Population: Who Do We Care About?

- We've decided on a concept and measurement strategy
- Now we need to decide who we're interested in
- This is the **target population**
- Often the target population is obvious based on research question
- Typical default choices for public opinion research:
  - Adult population
  - Adult citizen population
  - Registered voters
- Can get more vague too
  - Election forecasting: People who will vote in November
  - Market research: people in the market for a used car

From Population to Sampling Frame

- The **sampling frame** is the set of people who *could* be included in our survey
- Think of this as a list of actual people/units (we may not know anything about the people, but the list exists)
- We'd like this to perfectly reflect the population, but we might have **undercoverage** or **overcoverage** error
- Sampling frame will be highly affected by **survey mode** (phone, internet, in-person, etc.)

Ideology in Philly: A Sampling Frame

We are interested in the population of registered voters in Philadelphia. What are some potential sampling frames and potential coverage error?

Ideology in Philly: A Sampling Frame

We are interested in the population of registered voters in Philadelphia. What are some potential sampling frames and potential coverage error?

- Phone survey: People listed in voter file with phone numbers

Ideology in Philly: A Sampling Frame

We are interested in the population of registered voters in Philadelphia. What are some potential sampling frames and potential coverage error?

- Phone survey: People listed in voter file with phone numbers
- Internet survey: People listed in voter file with email addresses

Ideology in Philly: A Sampling Frame

We are interested in the population of registered voters in Philadelphia. What are some potential sampling frames and potential coverage error?

- Phone survey: People listed in voter file with phone numbers
- Internet survey: People listed in voter file with email addresses
- Phone survey: phone numbers in 215, 610, and 267 area codes

Ideology in Philly: A Sampling Frame

We are interested in the population of registered voters in Philadelphia. What are some potential sampling frames and potential coverage error?

- Phone survey: People listed in voter file with phone numbers
- Internet survey: People listed in voter file with email addresses
- Phone survey: phone numbers in 215, 610, and 267 area codes
- **Door-to-door survey: All people listed in voter file with adddress**

From Sampling Frame to Sample

Once we have a list of eligible units, we select from that sampling frame to generate a **sample** — people who actually participate in the survey

From Sampling Frame to Sample

Once we have a list of eligible units, we select from that sampling frame to generate a **sample** — people who actually participate in the survey

Broadly, two types of sampling: **probability** and **non-probability samples**

From Sampling Frame to Sample

Once we have a list of eligible units, we select from that sampling frame to generate a **sample** — people who actually participate in the survey

Broadly, two types of sampling: **probability** and **non-probability samples**

Probability samples: every unit in the sampling frame has a *known*, *nonzero* probability of being invited to participate in the survey

- generally preferable
- we'll focus on this type of sample for now

From Sampling Frame to Sample

Once we have a list of eligible units, we select from that sampling frame to generate a **sample** — people who actually participate in the survey

Broadly, two types of sampling: **probability** and **non-probability samples**

Probability samples: every unit in the sampling frame has a *known*, *nonzero* probability of being invited to participate in the survey

- generally preferable
- we'll focus on this type of sample for now

Non-probability sample: units have an *unknown* probability of being sampled

From Sampling Frame to Sample

For **probability samples**:

- We **randomly sample** members of the sampling frame to invite to participate in the survey
- Random sampling allows us to use probability theory to quantify the difference between a statistic in the sample and the population
- $\rightsquigarrow$ more on that next week

From Sampling Frame to Sample

Probability samples: Every member of the sampling frame has a known, nonzero probability of being invited to participate in the survey

From Sampling Frame to Sample

Probability samples: Every member of the sampling frame has a known, nonzero probability of being invited to participate in the survey

The probability depends on the sampling design

- Simple random sampling: give every unit an equal probability

From Sampling Frame to Sample

Probability samples: Every member of the sampling frame has a known, nonzero probability of being invited to participate in the survey

The probability depends on the sampling design

- Simple random sampling: give every unit an equal probability
- Cluster sampling: put units into groups, then sample groups as a whole

From Sampling Frame to Sample

Probability samples: Every member of the sampling frame has a known, nonzero probability of being invited to participate in the survey

The probability depends on the sampling design

- Simple random sampling: give every unit an equal probability
- Cluster sampling: put units into groups, then sample groups as a whole
- Stratified sampling: first sample groups, then sample units within groups

From Sampling Frame to Sample

Probability samples: Every member of the sampling frame has a known, nonzero probability of being invited to participate in the survey

The probability depends on the sampling design

- Simple random sampling: give every unit an equal probability
- Cluster sampling: put units into groups, then sample groups as a whole
- Stratified sampling: first sample groups, then sample units within groups
- Oversample: assign higher probability to some units than others

From Sampling Frame to Sample

Probability samples: Every member of the sampling frame has a known, nonzero probability of being invited to participate in the survey

The probability depends on the sampling design

- Simple random sampling: give every unit an equal probability
- Cluster sampling: put units into groups, then sample groups as a whole
- Stratified sampling: first sample groups, then sample units within groups
- Oversample: assign higher probability to some units than others

Sampling Error

Because we select a **sample**, the average value of the construct in our sample may differ from that of the population as a whole

- if you ask 1,000 people how tall they are, unlikely the average will be *exactly* the same as the average in the population

This is called **sampling error**

- relatively innocuous because we can quantify it
- as our sample size increases, this type of error goes to 0 very quickly

Ideology in Philly: Selecting a Sample

Suppose we select a random sample of individuals from the voter file, knock on their door and interview everyone who lives in their household.

What is the probability that any given individual is included in the sample?

Ideology in Philly: Selecting a Sample

Suppose we select a random sample of individuals from the voter file, knock on their door and interview everyone who lives in their household.

What is the probability that any given individual is included in the sample?

Ideology in Philly: Selecting a Sample

Suppose we select a random sample of individuals from the voter file, knock on their door and interview everyone who lives in their household.

What is the probability that any given individual is included in the sample?

Not everyone has an equal probability of selection!

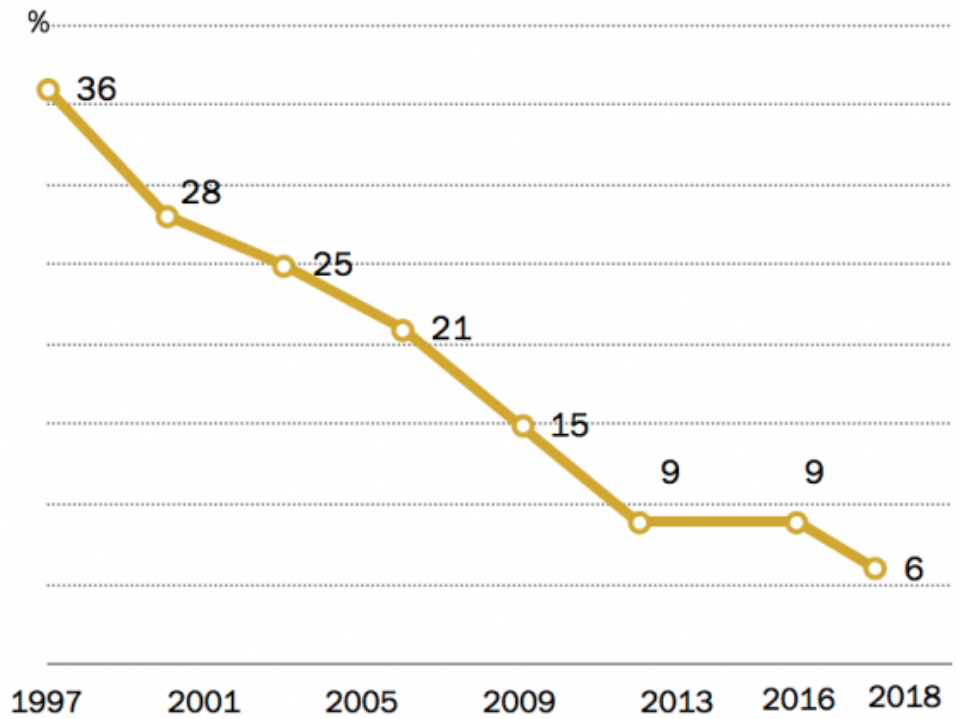
Instead, the probability an individual is sampled is proportional to the number of people in their household

- people in large households will be overrepresented

Intentions and Outcomes

After brief plateau, telephone survey response rates have fallen again

Response rate by year (%)



Note: Response rate is AAPOR RR3. Only landlines sampled 1997-2006. Rates are typical for surveys conducted in each year.
Source: Pew Research Center telephone surveys conducted 1997-2018.

PEW RESEARCH CENTER

From Intended Sample to Realized Sample

- We may invite a random sample to participate, but those who actually are not random
- Two types of **nonresponse** error:
  - **unit nonresponse**: entire sampled units do not complete the survey
  - **item nonresponse**: individuals may not respond to individual survey questions
- Nonresponse is a problem if those who respond are systematically different from those who do not. Can lead to **nonresponse bias**

Ideology in Philly: Nonresponse Bias

In our door-to-door survey of Philly registered voters, what might be some potential concerns with nonresponse?

Opt-In Surveys: No Well-Defined Sampling Frame

- Opt-in surveys (incl. most internet surveys) are **non-probability samples**
- No defined sampling frame and unknown inclusion probabilities
- Makes the statistical theory much more difficult — a big cause of concern 10-15 years ago
- Now there's recognition that no sample is really random due to nonresponse anyway

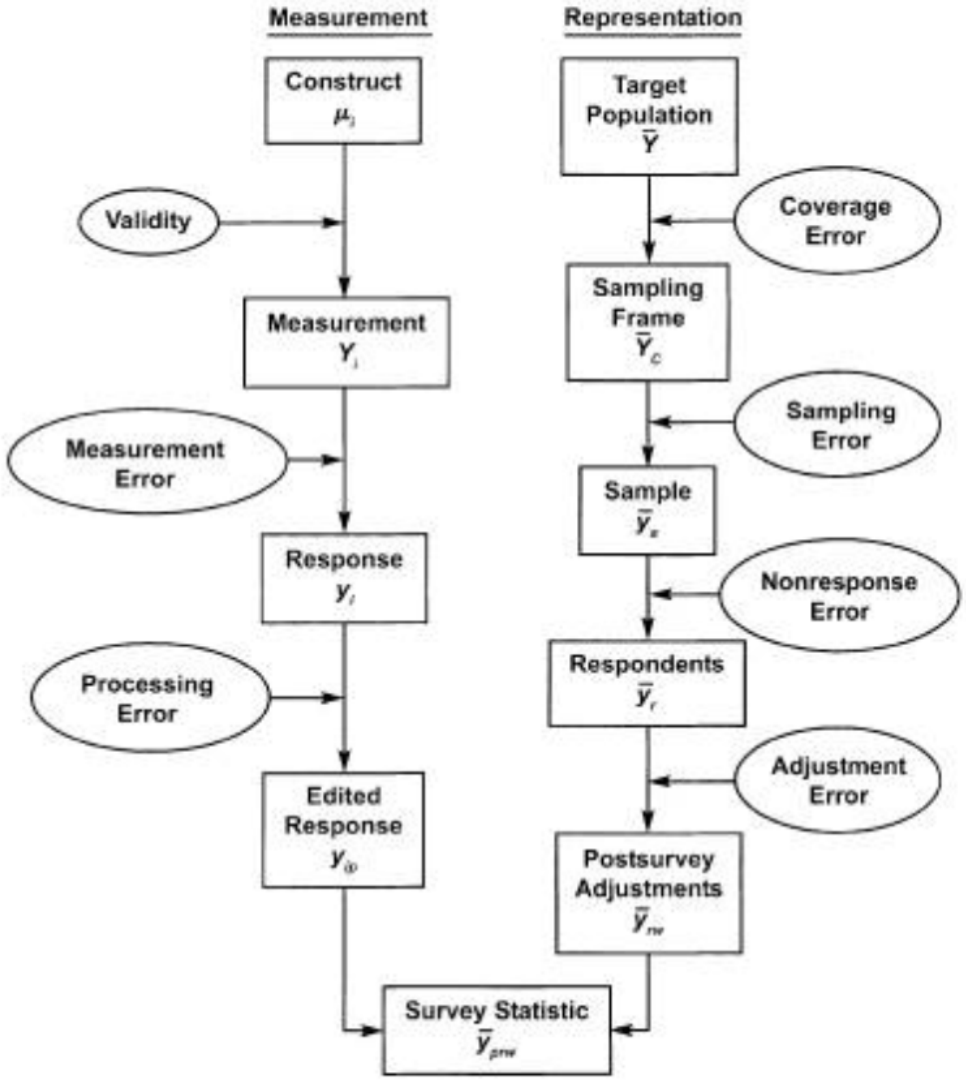


Figure 2.5 Survey life cycle from a quality perspective.

Application of Survey Error: Lee and Zhang (2017)

Application of Survey Error: Lee and Zhang (2017)

Research question: How does legibility affect state capacity?

- Theory: states need *systematic*, *standardized* measures about citizens
- Example: without information on economic activity, can't enforce tax laws
- Empirical problem: how should we measure the extent of legibility across time and space?

Application of Survey Error: Lee and Zhang (2017)

Research question: How does legibility affect state capacity?

- Theory: states need *systematic*, *standardized* measures about citizens
- Example: without information on economic activity, can't enforce tax laws
- Empirical problem: how should we measure the extent of legibility across time and space?

Proposed measurement: Quality of age statistics in Census

Application of Survey Error: Lee and Zhang (2017)

Research question: How does legibility affect state capacity?

- Theory: states need *systematic*, *standardized* measures about citizens
- Example: without information on economic activity, can't enforce tax laws
- Empirical problem: how should we measure the extent of legibility across time and space?

Proposed measurement: Quality of age statistics in Census

- Idea: when there is standardized knowledge, Census records on age are more likely to be accurate

Application of Survey Error: Lee and Zhang (2017)

Research question: How does legibility affect state capacity?

- Theory: states need *systematic*, *standardized* measures about citizens
- Example: without information on economic activity, can't enforce tax laws
- Empirical problem: how should we measure the extent of legibility across time and space?

Proposed measurement: Quality of age statistics in Census

- Idea: when there is standardized knowledge, Census records on age are more likely to be accurate
- But how do we know if Census records are accurate?

Application of Survey Error: Lee and Zhang (2017)

Research question: How does legibility affect state capacity?

- Theory: states need *systematic*, *standardized* measures about citizens
- Example: without information on economic activity, can't enforce tax laws
- Empirical problem: how should we measure the extent of legibility across time and space?

Proposed measurement: Quality of age statistics in Census

- Idea: when there is standardized knowledge, Census records on age are more likely to be accurate
- But how do we know if Census records are accurate?
- When Census records are inaccurate it tends to manifest in "clumping"

Example from 1971 Moroccan Census

Enumerator: What is your age?

Respondent: Who me? Our generation was unrecorded. We didn't have any. No date of birth. Nothing.

Enumerator: How many (years), how many? Estimate.

Respondent: How am I going to estimate? I have nothing to estimate with. I can tell you that I am 60 years; 70 I haven't reached.

—from Quandt (1973, 45), cited in Lee and Zhang (2017)

Example from 1971 Moroccan Census

Enumerator: What is your age?

Respondent: Who me? Our generation was unrecorded. We didn't have any. No date of birth. Nothing.

Enumerator: How many (years), how many? Estimate.

Respondent: How am I going to estimate? I have nothing to estimate with. I can tell you that I am 60 years; 70 I haven't reached.

—from Quandt (1973, 45), cited in Lee and Zhang (2017)

What type of survey error is this?

“Lumping” in the Age Distribution

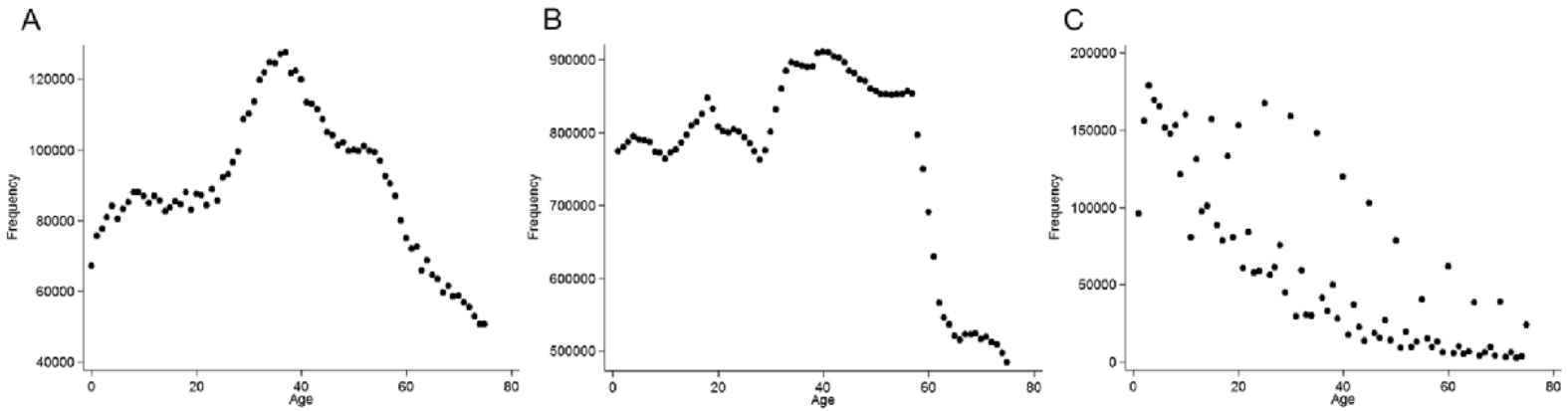


Figure 1. The effect of demographic shocks on the smoothness of age curves: *A*, Switzerland, 2000; *B*, France, 2006; *C*, Sierra Leone, 2004

Lee and Zhang (2017) Statistic

- To quantify the extent of lumping, first note that we might expect all terminal digits to be equally likely (e.g., 2$\underline{2}$, 2$\underline{3}$, 2$\underline{4}$, 2$\underline{5}$)* or **uniformly distributed**

Lee and Zhang (2017) Statistic

- To quantify the extent of lumping, first note that we might expect all terminal digits to be equally likely (e.g., 2$\underline{2}$, 2$\underline{3}$, 2$\underline{4}$, 2$\underline{5}$)* or **uniformly distributed**
- Calculate deviation from this "equally likely" baseline

Lee and Zhang (2017) Statistic

- To quantify the extent of lumping, first note that we might expect all terminal digits to be equally likely (e.g., 22, 23, 24, 25)* or **uniformly distributed**
- Calculate deviation from this "equally likely" baseline
- Lower deviation = more legibility (so the argument goes)

Lee and Zhang (2017) Statistic

- To quantify the extent of lumping, first note that we might expect all terminal digits to be equally likely (e.g., 22, 23, 24, 25)* or **uniformly distributed**
- Calculate deviation from this "equally likely" baseline
- Lower deviation = more legibility (so the argument goes)
- Consider: Is this a valid measure of the construct they're interested in?

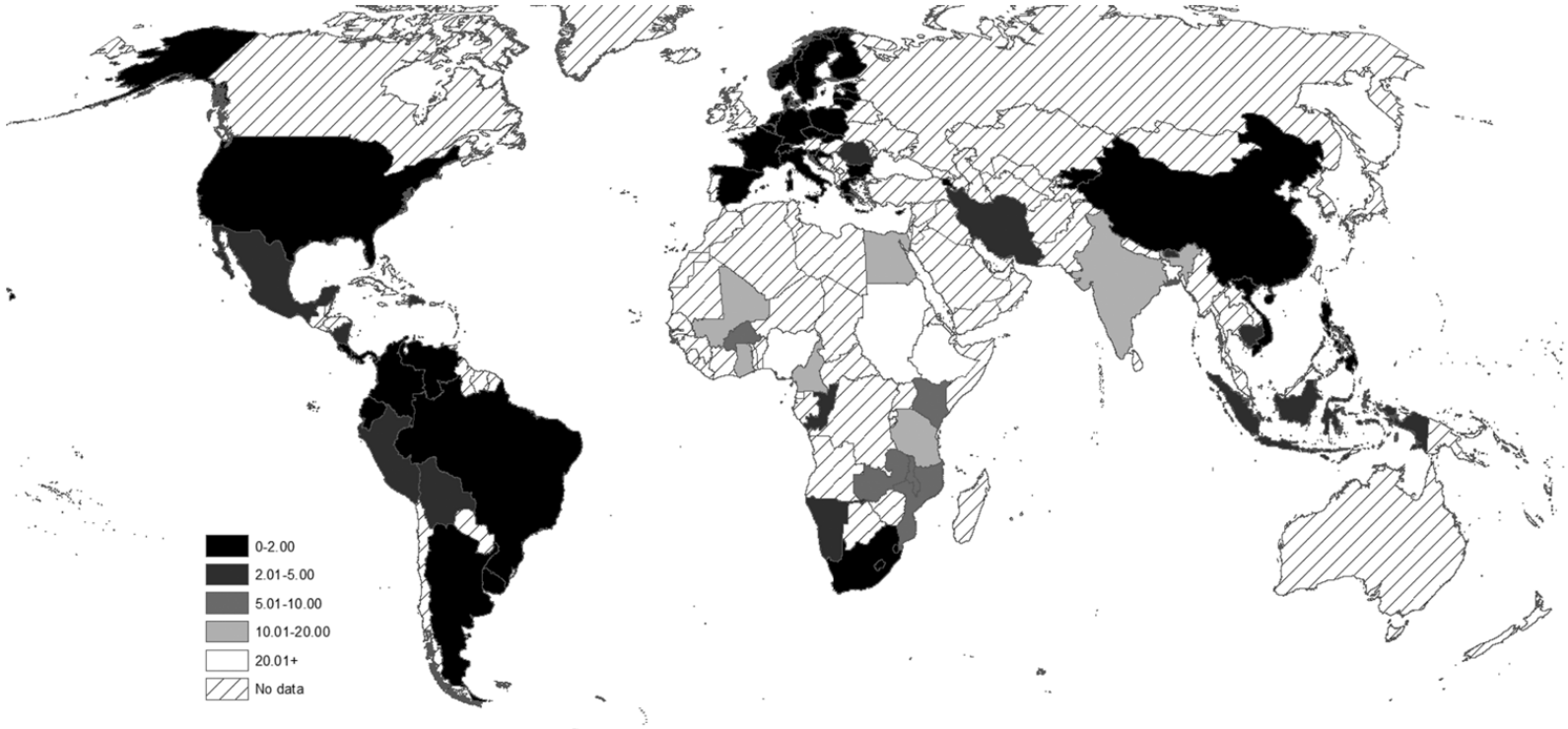


Figure 3. Myers scores by country, 2005–12

Results

Table 6. Legibility and Public Goods: National-Level Results

	Mortality (1)	Mortality (2)	Literacy (3)	Literacy (4)	Enrollment (5)	Enrollment (6)
Legibility	−.663**	−.283**	.797**	.507**	.586**	.229*
	(.0398)	(.0481)	(.0649)	(.0839)	(.0867)	(.0942)
GDP per capita		−.531**		.355**		.367**
		(.0572)		(.0738)		(.0955)
Democracy		−.0835*		.104		.0736
		(.0364)		(.0726)		(.0636)
Population density		−.146**		.0374		.131+
		(.0322)		(.0456)		(.0678)
Terrain ruggedness		.0551+		.113+		.184**
		(.0281)		(.0581)		(.0655)
Constant	.333**	.221**	−.597**	−.514**	−.426*	−.413**
	(.0687)	(.0531)	(.212)	(.193)	(.163)	(.147)
Number of observations	326	326	188	188	244	244
Number of countries	111	111	84	84	105	105
R^2	.744	.888	.673	.758	.445	.576

Note. Decade-specific intercepts are suppressed. Standard errors are in parentheses and are clustered by country.

+ $p < .10$.

* $p < .05$.

** $p < .01$.

Examining Design Features of the American National Election Studies

ANES Design Features

Get some practice identifying key design features. Download the guide for the 2020 version of the American National Election Study from Canvas ("Misc" folder). Read the section "Sample Design and Respondent Recruitment."

- What is the target population?
- What is the sampling frame(s)?
- How were respondents recruited?
- What are the interview modes (i.e. how were respondents interviewed)?
- What is the sample size?
- Other design features?

ANES Design Features

Population:

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered."

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment:

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment: (1) "the fresh cross-sectional group received mail invitations.";

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment: (1) "the fresh cross-sectional group received mail invitations.";(2) "The 2016 ANES respondents were invited by email where possible, with letters used if there was no email on file or after an initial non-response"

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment: (1) "the fresh cross-sectional group received mail invitations.";(2) "The 2016 ANES respondents were invited by email where possible, with letters used if there was no email on file or after an initial non-response"

Mode:

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment: (1) "the fresh cross-sectional group received mail invitations.";(2) "The 2016 ANES respondents were invited by email where possible, with letters used if there was no email on file or after an initial non-response"

Mode:(1) web only; (2) mixed web-phone; (3) mixed video-web-phone

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment: (1) "the fresh cross-sectional group received mail invitations.";(2) "The 2016 ANES respondents were invited by email where possible, with letters used if there was no email on file or after an initial non-response"

Mode:(1) web only; (2) mixed web-phone; (3) mixed video-web-phone

Sample size:

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment: (1) "the fresh cross-sectional group received mail invitations.";(2) "The 2016 ANES respondents were invited by email where possible, with letters used if there was no email on file or after an initial non-response"

Mode:(1) web only; (2) mixed web-phone; (3) mixed video-web-phone

Sample size: total 5,441 in fresh cross-section

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment: (1) "the fresh cross-sectional group received mail invitations.";(2) "The 2016 ANES respondents were invited by email where possible, with letters used if there was no email on file or after an initial non-response"

Mode:(1) web only; (2) mixed web-phone; (3) mixed video-web-phone

Sample size: total 5,441 in fresh cross-section

Other design features:

ANES Design Features

Population: "The target population for the fresh cross-section was the 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia."

Sampling frame:(1) "The sampling frame for the fresh cross-section was lists of residential addresses where mail is delivered." (2) "The 2016 ANES group consisted of respondents who had participated in the ANES 2016 Time Series Study."

Recruitment: (1) "the fresh cross-sectional group received mail invitations.";(2) "The 2016 ANES respondents were invited by email where possible, with letters used if there was no email on file or after an initial non-response"

Mode:(1) web only; (2) mixed web-phone; (3) mixed video-web-phone

Sample size: total 5,441 in fresh cross-section

Other design features: incentives, re-contact, links to other data sets