J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

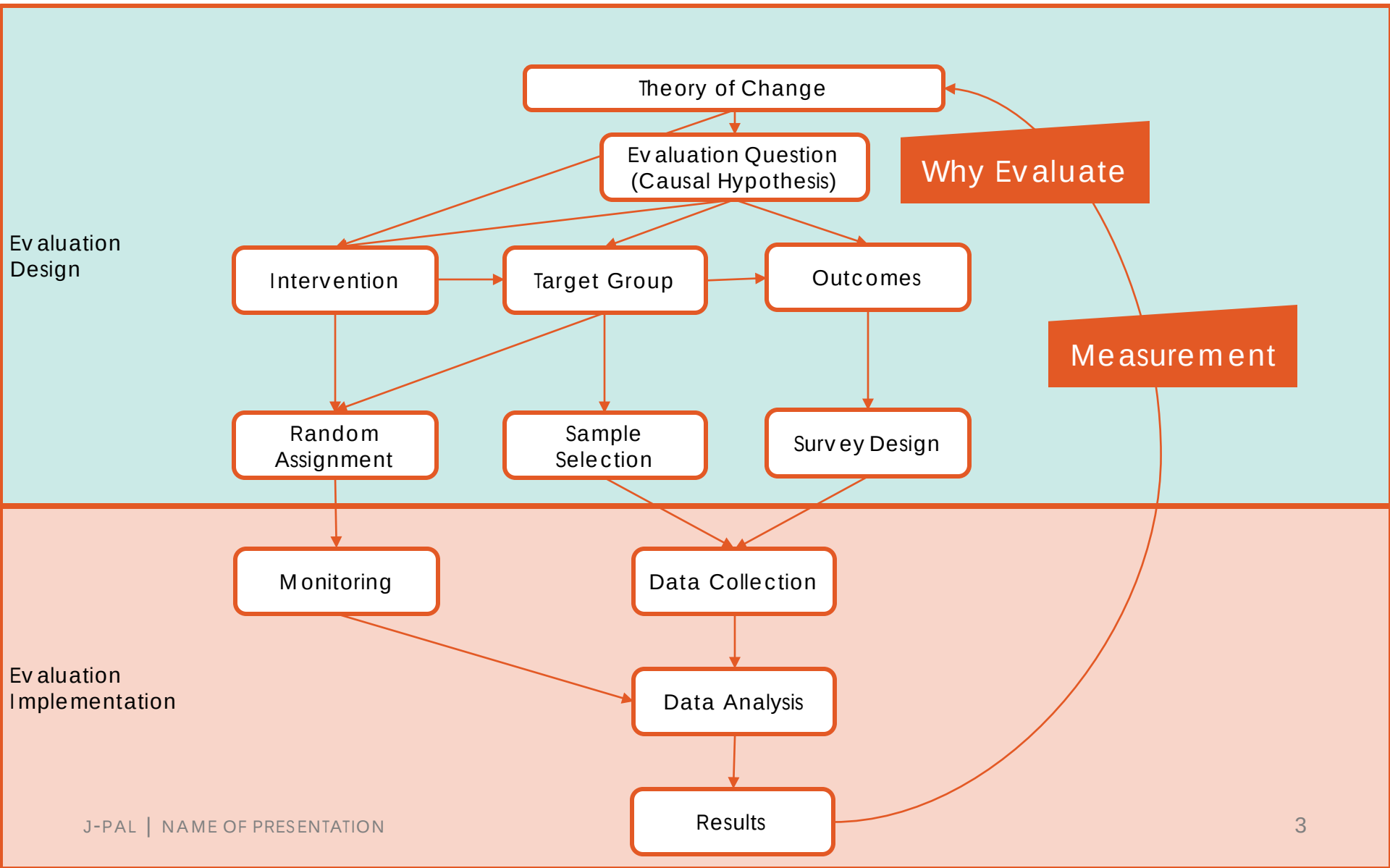
Measurement

Outcomes, Indicators, Data

Course Overview

1. What is Evaluation?
2. **Measurement & Indicators**
3. Why Randomize?
4. How to Randomize?
5. Sampling and Sample Size
6. Threats and Analysis
7. Generalizability
8. Project from Start to Finish

Randomized Evaluation Process



J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Measurement

Kelsey Jack
Assistant Professor
Tufts University

Lecture Overview

1. **What to Measure**
   - Theory of Change, Outcomes, Indicators
2. **How to Measure It (Well)**
   - Sources of Measurement
   - Measurement Concepts
   - Response Process
   - Measurement Error
   - Best Practices

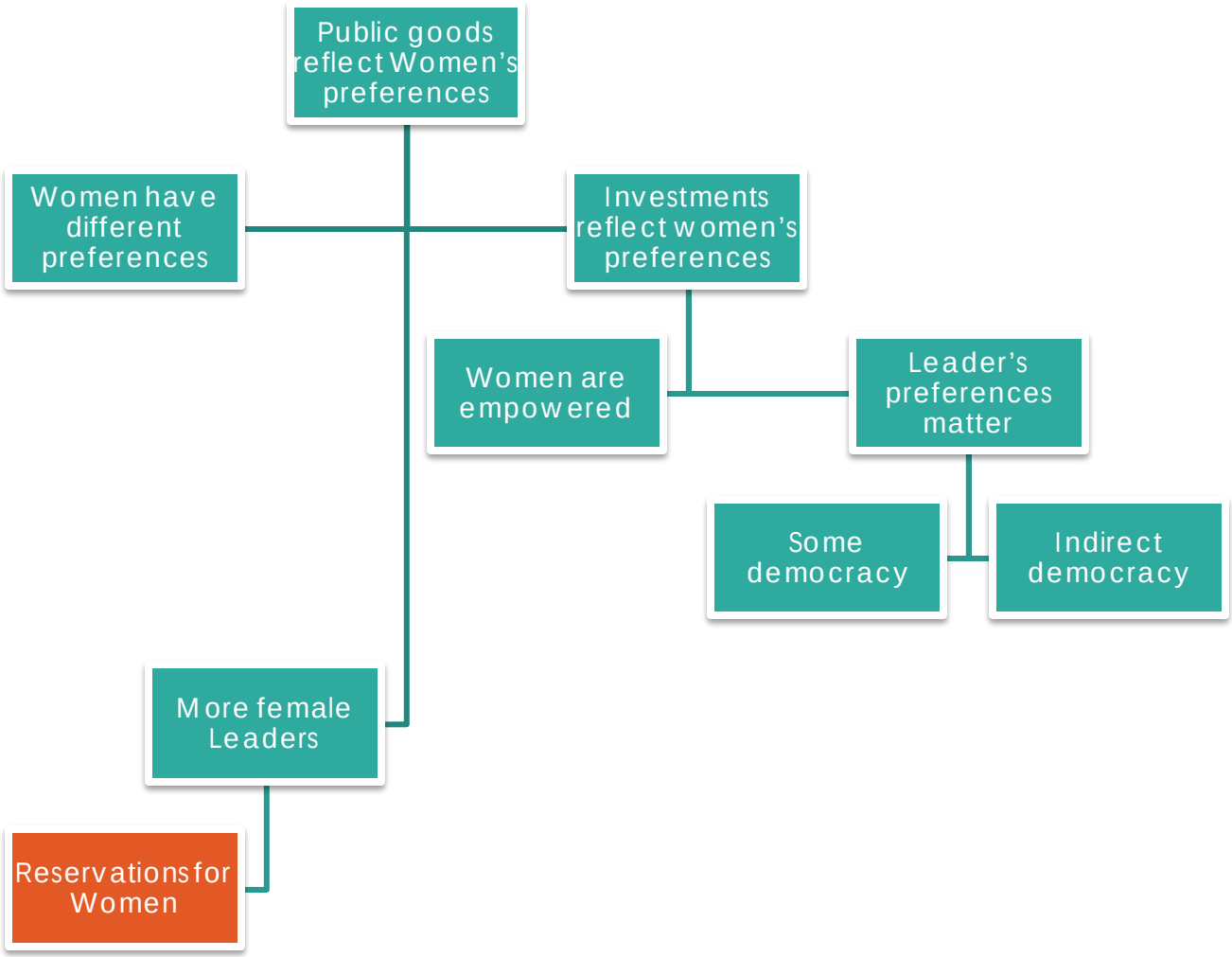
Lecture Overview

1. **What to Measure**
   - Theory of Change, Outcomes, Indicators
2. **How to Measure It (Well)**
   - Sources of Measurement
   - Measurement Concepts
   - Response Process
   - Measurement Error
   - Best Practices

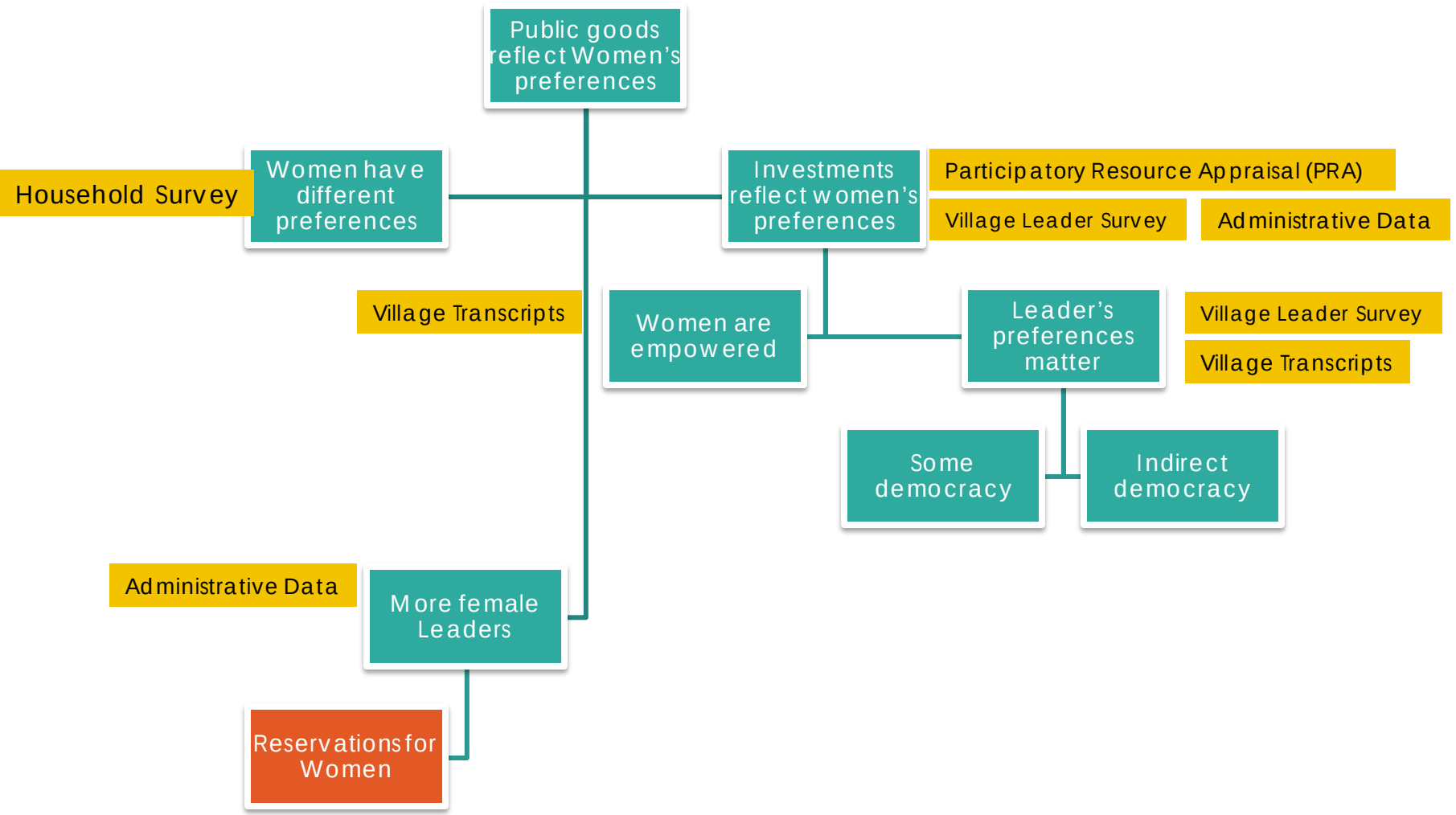
What to Measure

Women as Policymakers

Theory of Change



Theory of Change: How to measure?



Log Frame

	Objectives Hierarchy	Indicators	Sources of Verification	Assumptions / Threats
Impact (Goal/ Overall objective)	Public good investment represents women's preferences	Government spending	Administrative data: Budgets, Balance Sheets	Pradhan preferences matter: imperfect/some democracy
Outcome (Project Objective)	Women voice political views	Number of times a woman spoke	Transcript from village meeting	Women develop independent views
Outputs	More female Pradhans	Whether or not a Panchayat had a female Pradhan	Administrative records	The law is implemented, there is no backlash
Inputs (Activities)	Reservations for women	Law is passed	The constitution	The government realizes the need for women representation

Source: Roduner, Schlappi (2008) Logical Framework Approach and Outcome Mapping, A constructive Attempt of Synthesis,

Results, By State, By Issue

		West Bengal			Rajasthan		
Issue	**Investment Indicator**	**Issue Priority for**		**Investment Measure in Quota Villages**	**Issue Priority for**		**Investment Measure in Quota Villages**
		W	**M**		**W**	**M**	
Drinking Water	# facilities	**31%**	17%	9.09*	**54%**	49%	2.62*
Road Improvement	Road Condition (0-1)	**31%**	25%	0.18*	13%	**23%**	-0.08*
Irrigation	# facilities	4%	**20%**	-0.38	2%	4%	-0.02
Education	Informal education center	6%	**12%**	-0.06	5%	**13%**	-

How to Measure

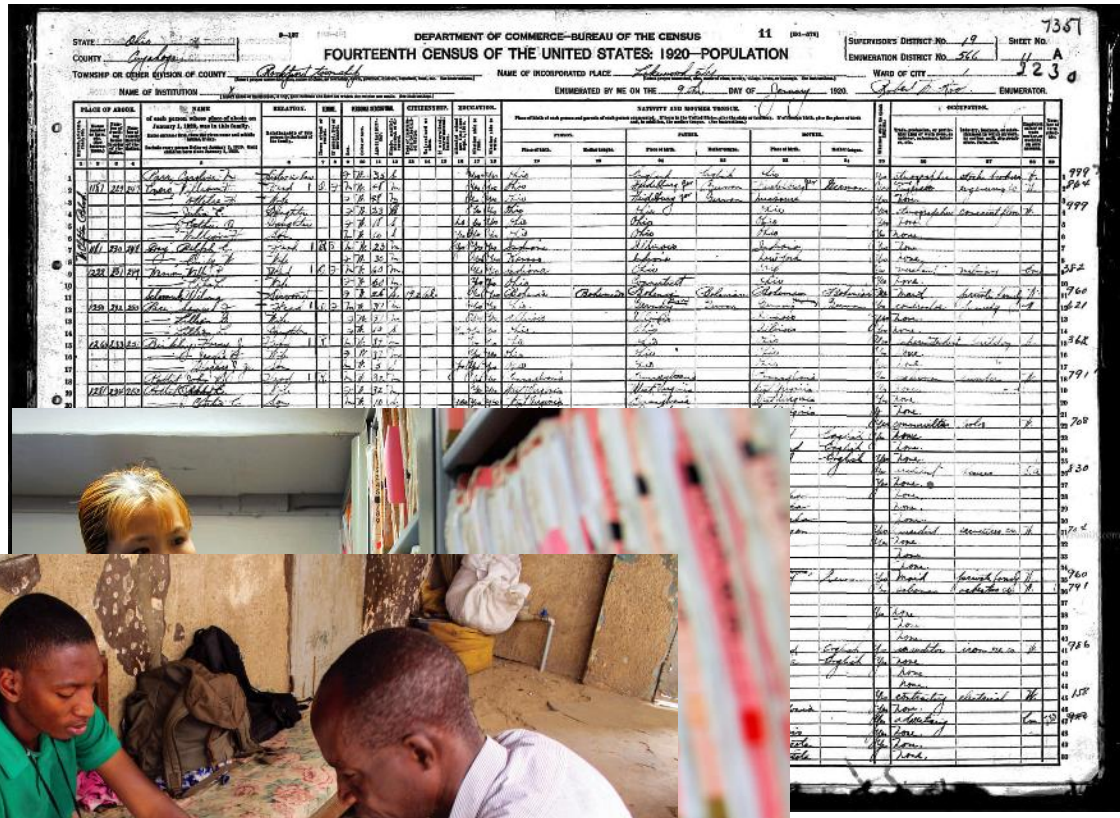
Sources of Measurement

First-order questions in measurement

- What data do you collect?
- Where do you get it?
- When do you get it?

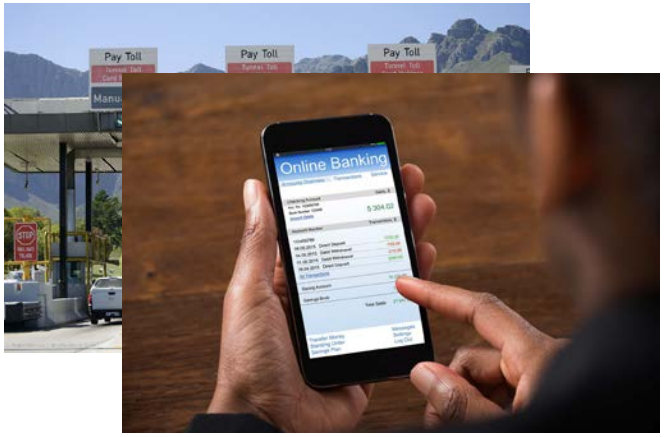
Where can we get data?

- Obtained from other sources
  - Publically available
  - Administrative data
  - Other secondary data
- Collected by researchers
  - Primary data



https://commons.wikimedia.org/wiki/File:Cuyahoga_County_US_Census_Form-Herbert_Birch_Kingston_1920.jpg
https://commons.wikimedia.org/wiki/File:US_Navy_090123-N-9760Z-004_Hospital_Corpsman_2nd_Class_Jennifer_Ross_files_medical_records_aboard_the_aircraft_carrier_USS_Nimitz_(CVN_68).jpg

Types and Sources of Data

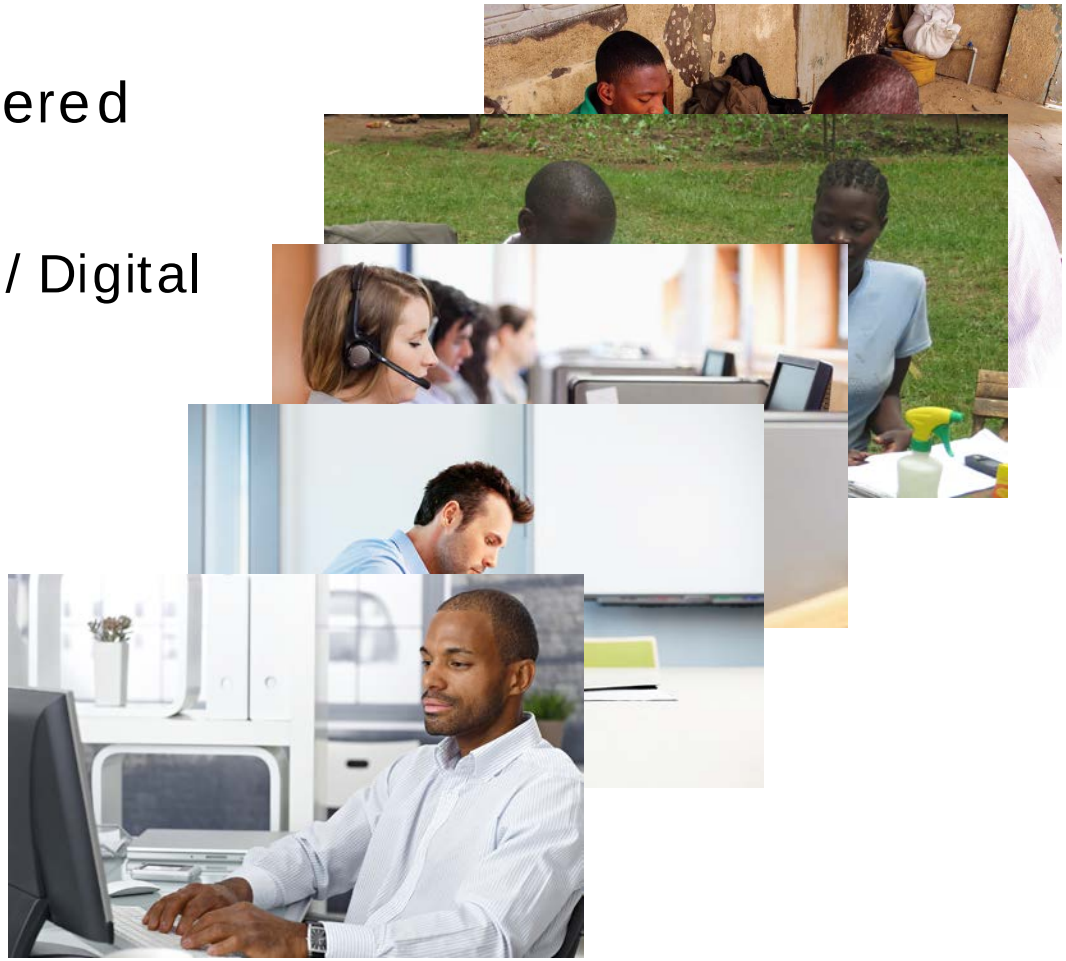
	Information about a person/ household / possessions	NOT about a person/ household / possessions
Information provided by a person		
Automatically generated		

Data collection on people

- Surveys
- Exams, tests, etc.
- Games
- Vignettes
- Direct Observation
- Diaries/Logs
- Focus groups
- Interviews

Survey: Modes of Data Collection

- Interviewer administered
  - Paper-based
  - Computer-assisted/ Digital
  - Telephone-based
- Self-administered
  - Paper
  - Computer/Digital



When to collect data

- Baseline
- During the intervention
  - Process, Monitoring of intervention
- Endline
- Follow-up
- Scale-up
- Intervention: M&E

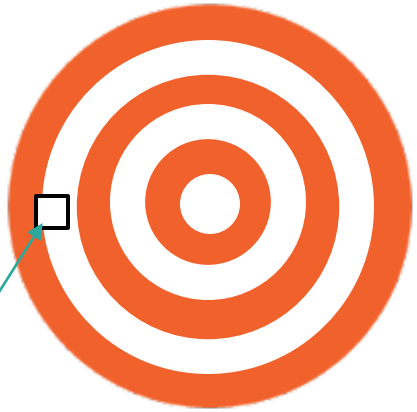
Ethics

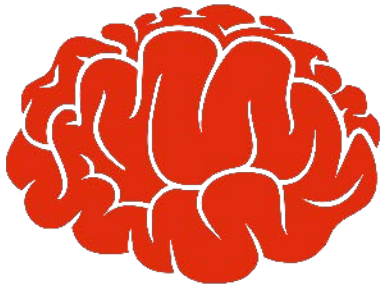
- "Experimenting on people"
- Belmont Principles
  - Respect for persons
  - Beneficence
  - Justice
- Institutional Review Boards (IRBs)

How to Measure

Concept

Concept of measurement

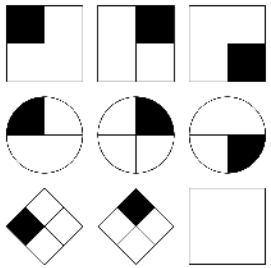




Data (Test Result)

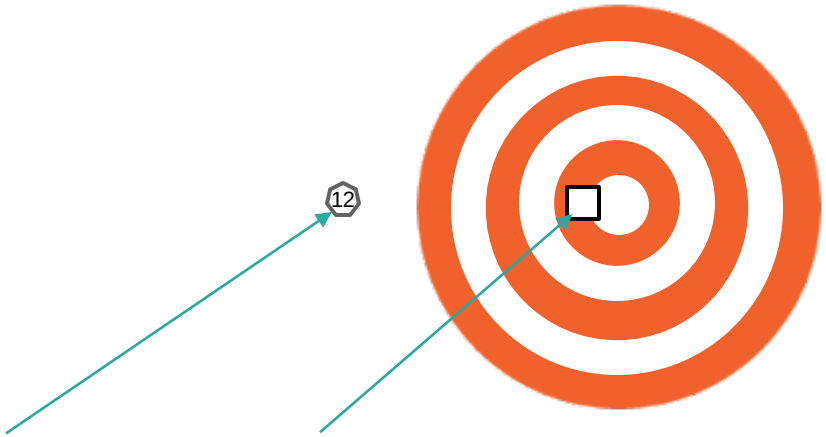
Indicator (IQ Test)

Construct (Intelligence)

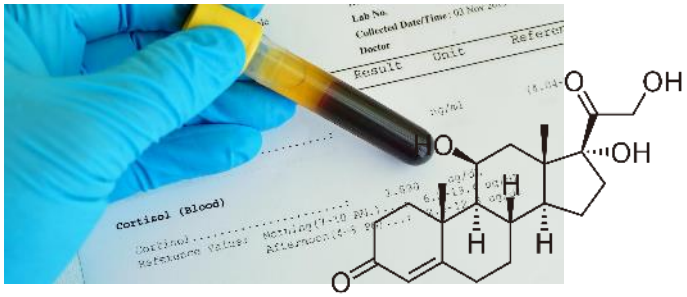


https://commons.wikimedia.org/wiki/File:Red_Silhouette_-_Brain.svg

Concept of measurement

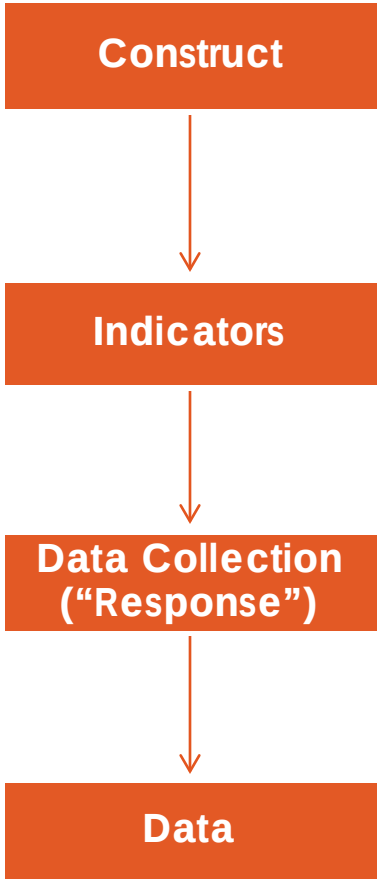


Data (Test Result) Indicator (Cortisol level) Construct (Stress)



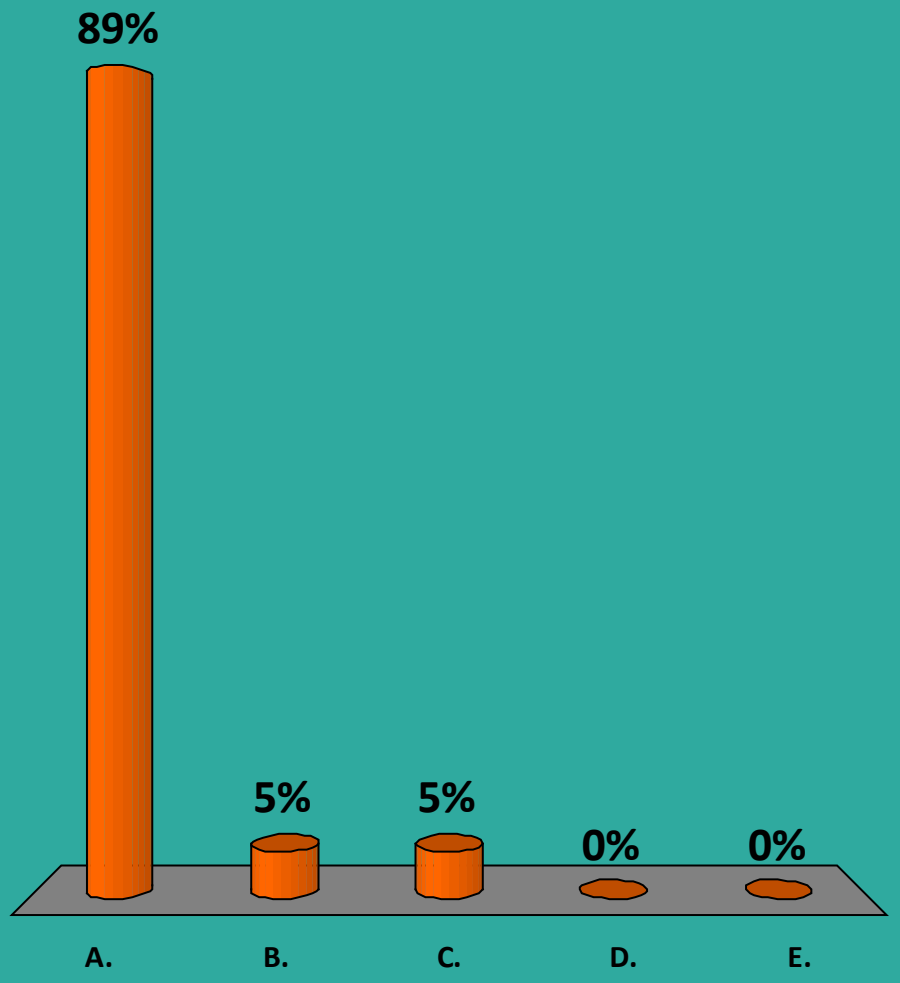
https://pixabay.com/en/despair-stress-alone-being-alone-862349/

Concept of measurement



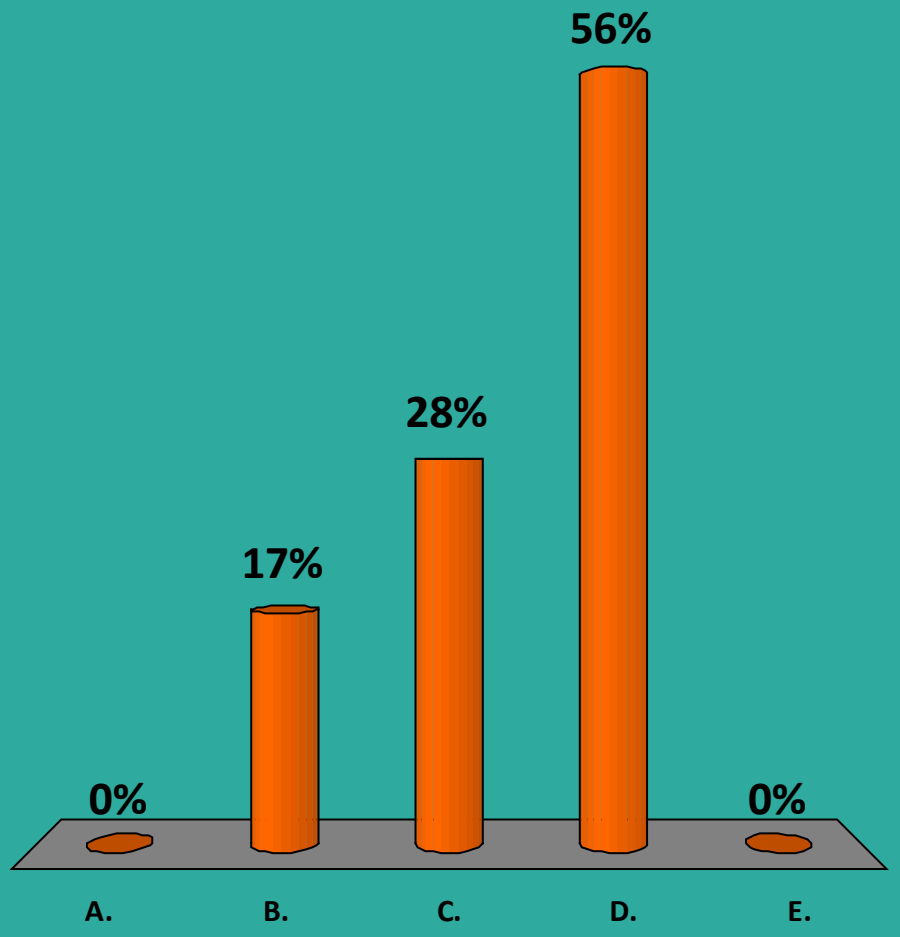
Empowerment is:

A. A construct
B. An indicator
C. A response
D. Data
E. Don't know



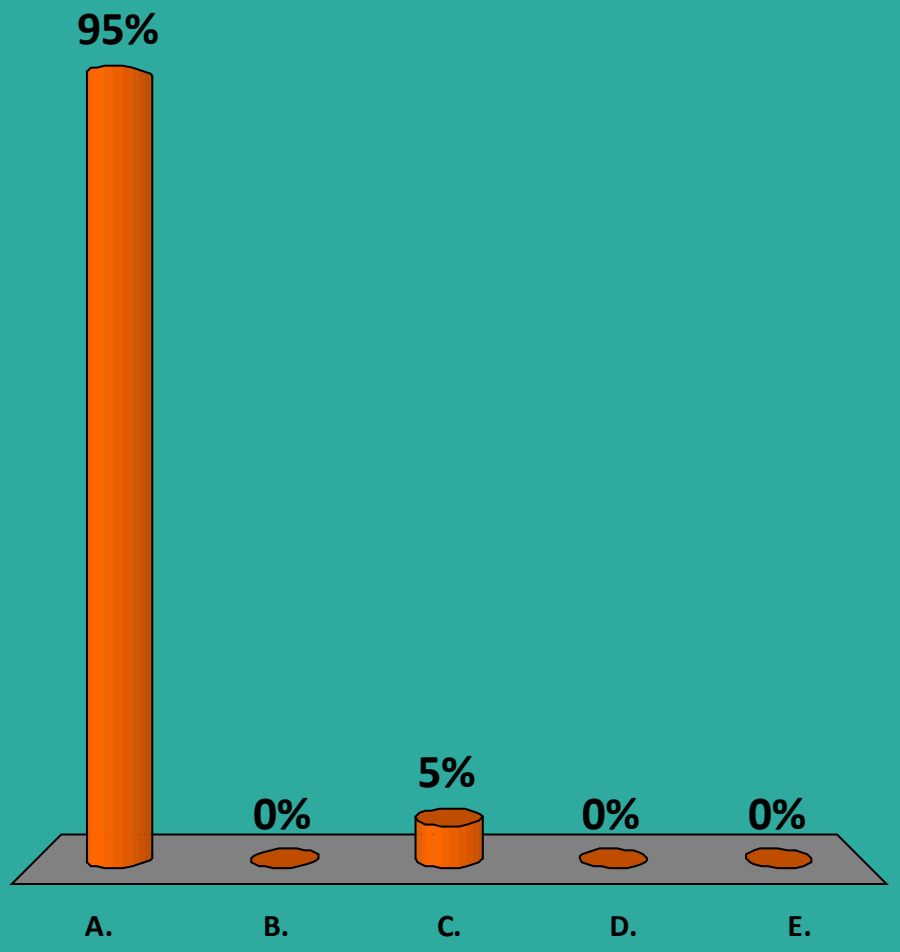
“Blood Pressure = 110/71 mm Hg” is:

A. A construct
B. An indicator
C. A response
D. Data
E. Don’t know



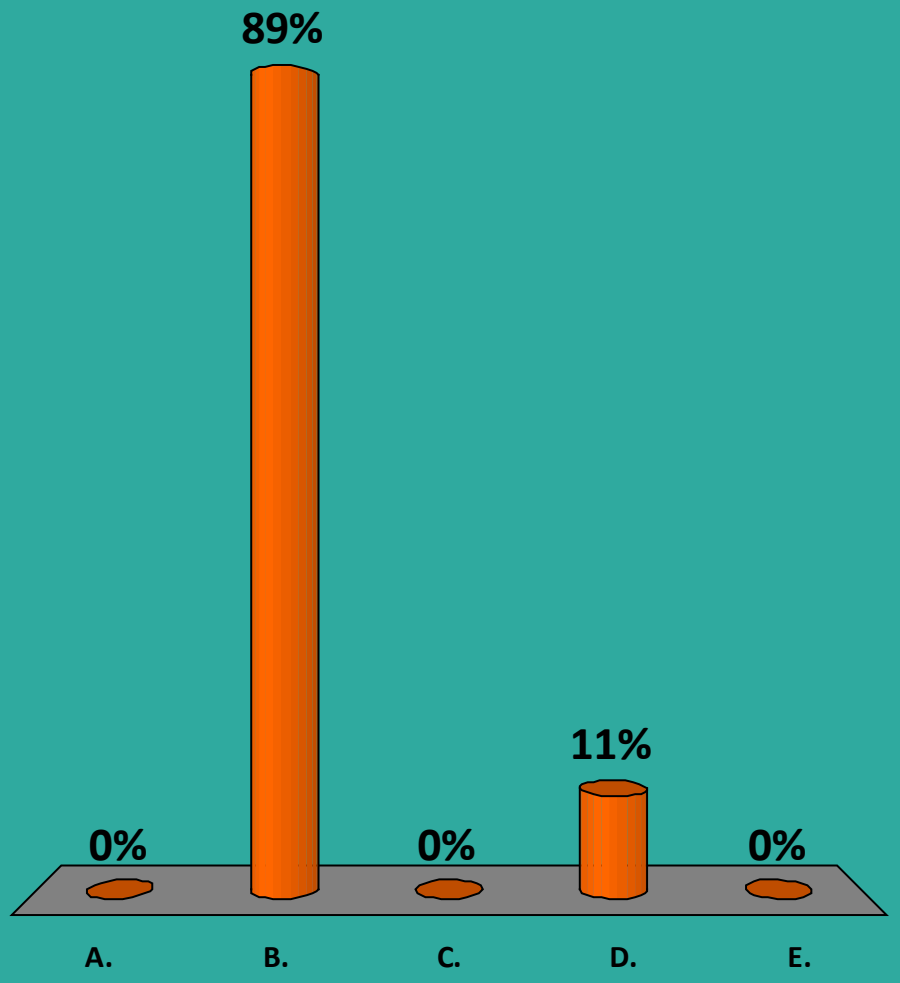
Discrimination is:

A. A construct
B. An indicator
C. A response
D. Data
E. Don't know



Kilograms of rice per hectare:

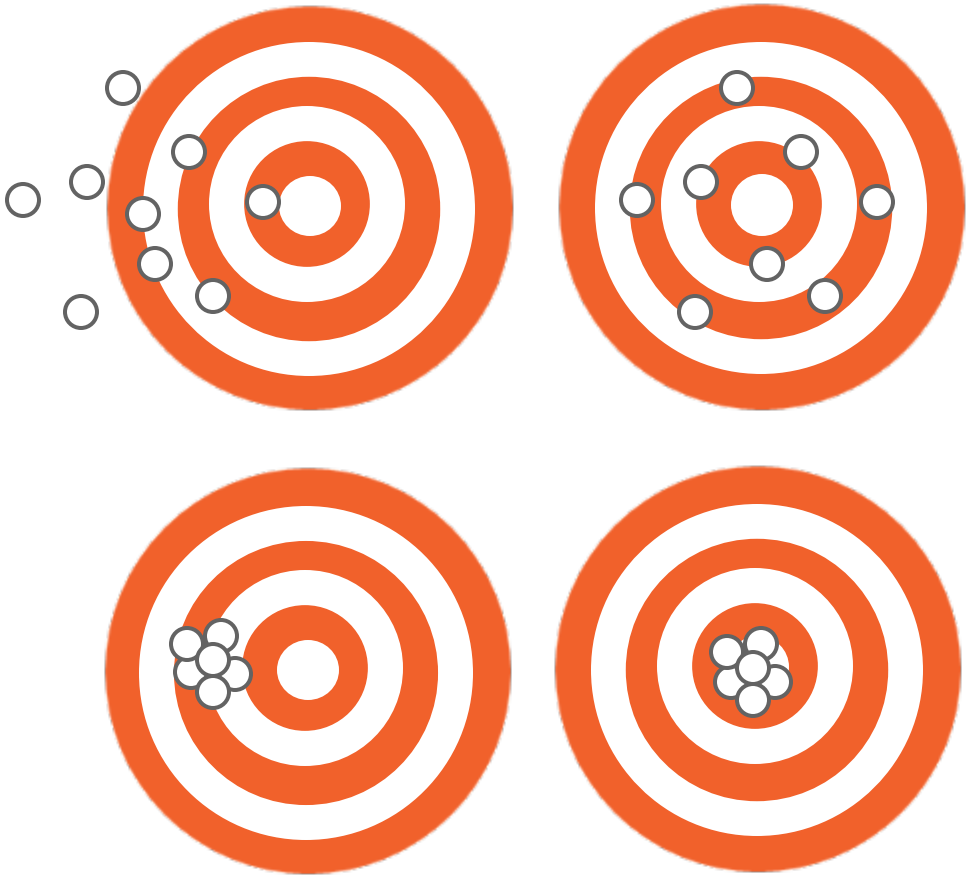
A. A construct
B. An indicator
C. A response
D. Data
E. Don't know



The goals of measurement

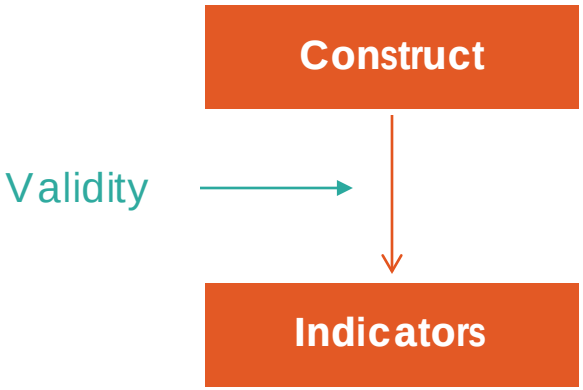
- Accuracy
- Unbiasedness
- Validity

- Precision
- Reliability



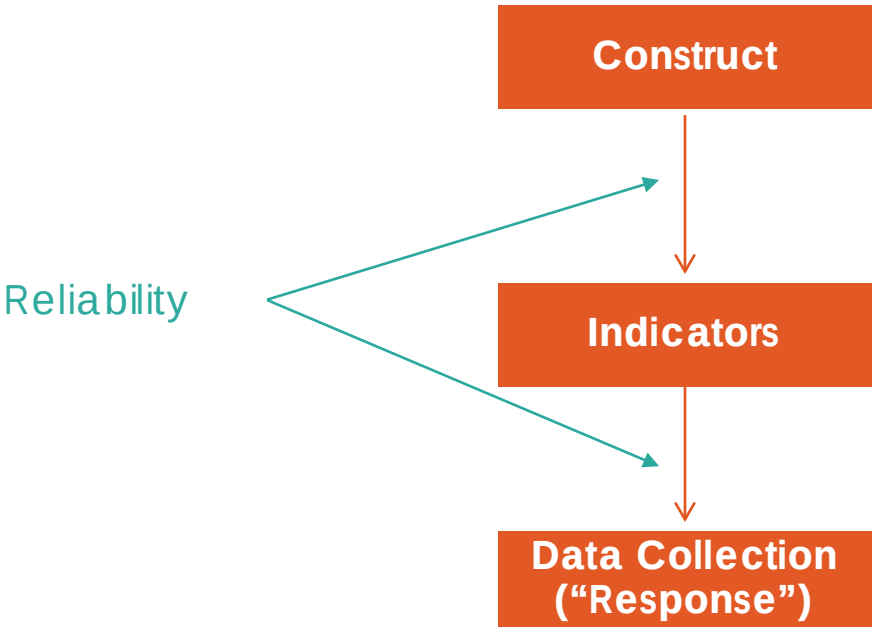
Validity

- In theory:
  - How well does the indicator map to the outcome? (e.g. IQ tests → intelligence)



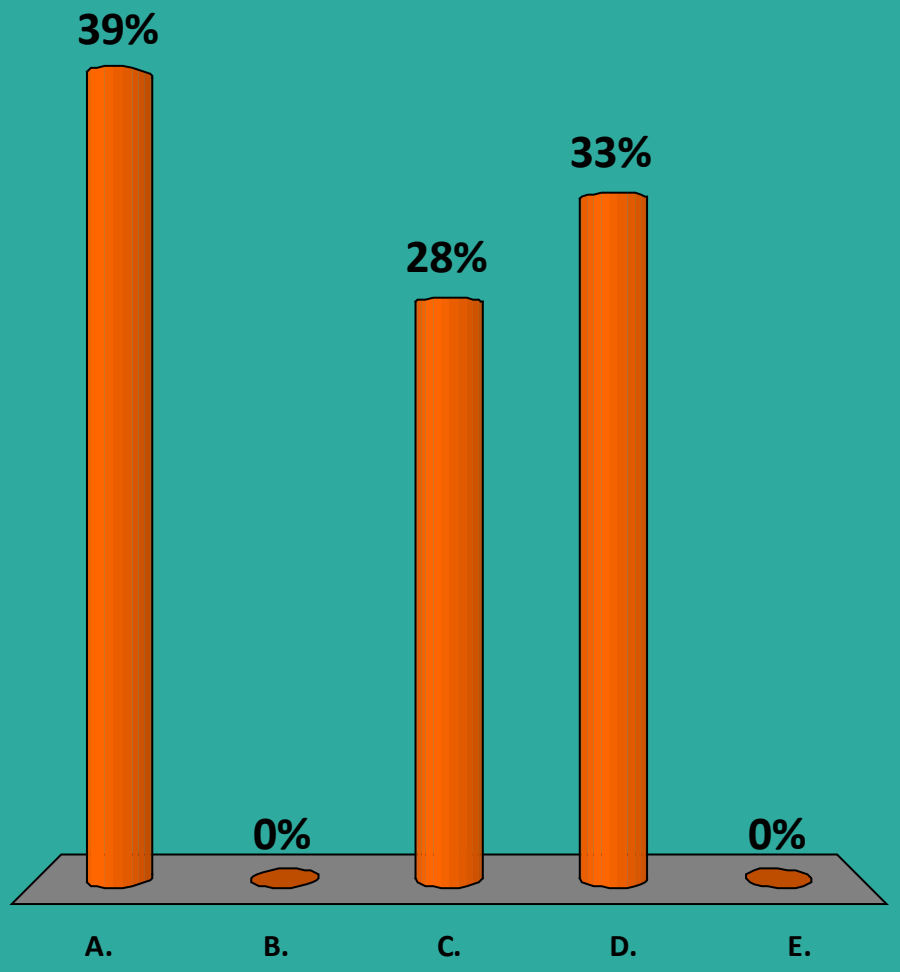
Reliability

- In theory:
  - The measure is consistent and precise vs. "noisy"



Which is worse?

A. Poor Validity
B. Poor Reliability
C. Equally bad
D. Depends
E. Don't know/can't say

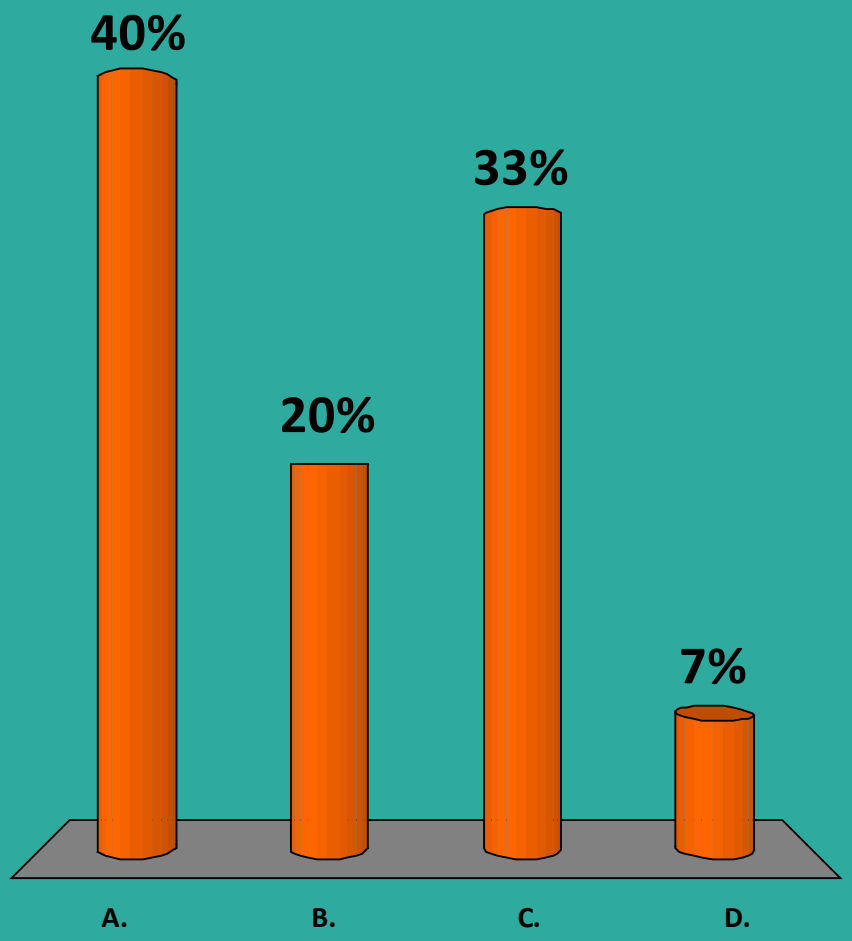


The problem

- With the following questions…

Outcome: annual consumption
Indicator: food expenditure in last week

A. Validity
B. Reliability
C. Both
D. Neither



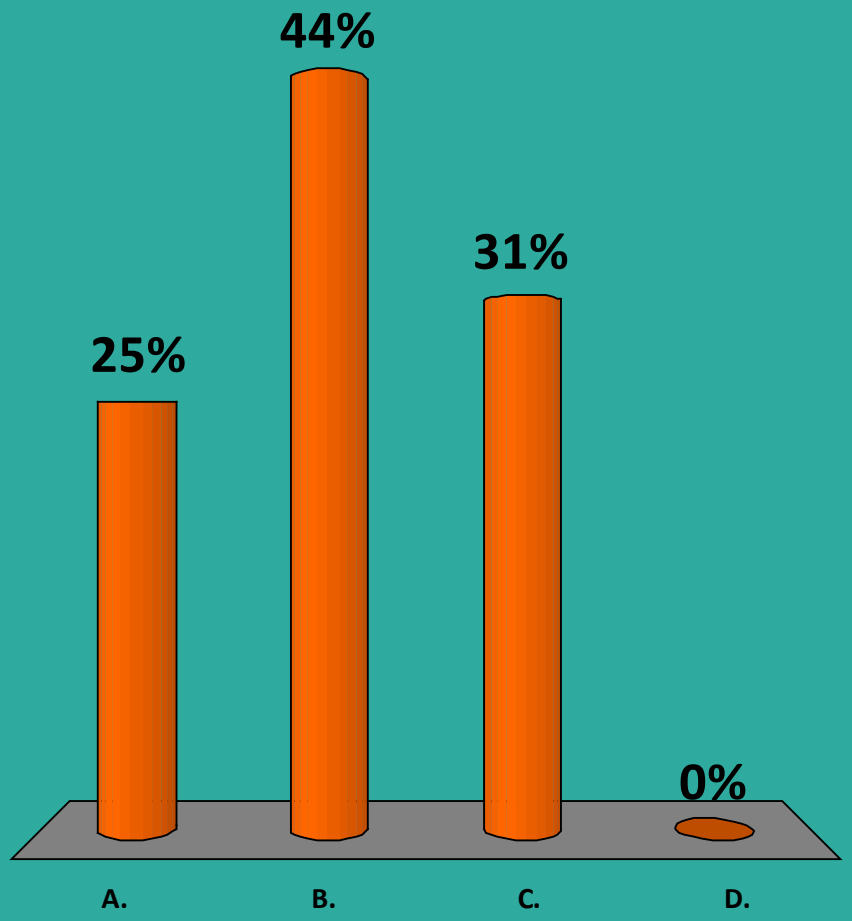
Outcome: annual consumption
Indicator: food expenditure in last three months

A. Validity

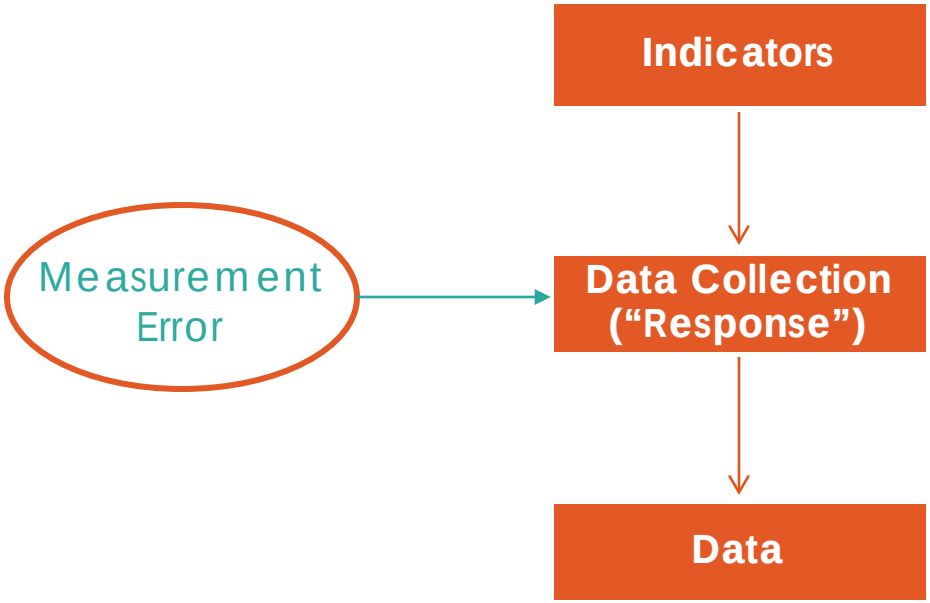
B. Reliability

C. Both

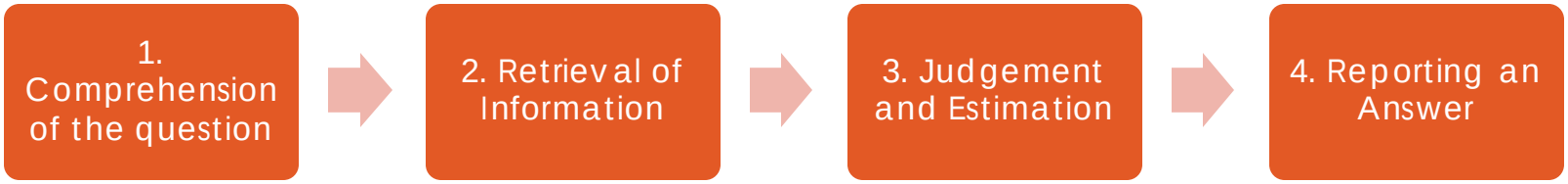
D. Neither



The Response Process



4-step Response Process



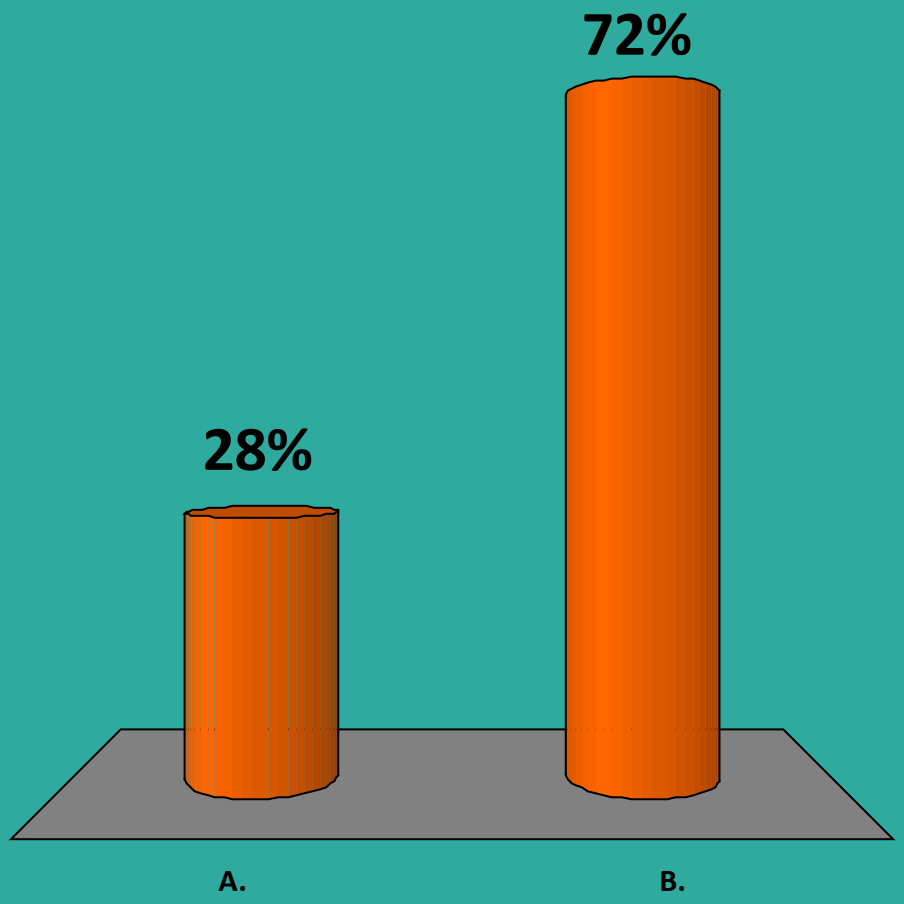
1. Comprehension of the question → 2. Retrieval of Information → 3. Judgement and Estimation → 4. Reporting an Answer

Step 1: Comprehension

Do you prefer sitting or not?

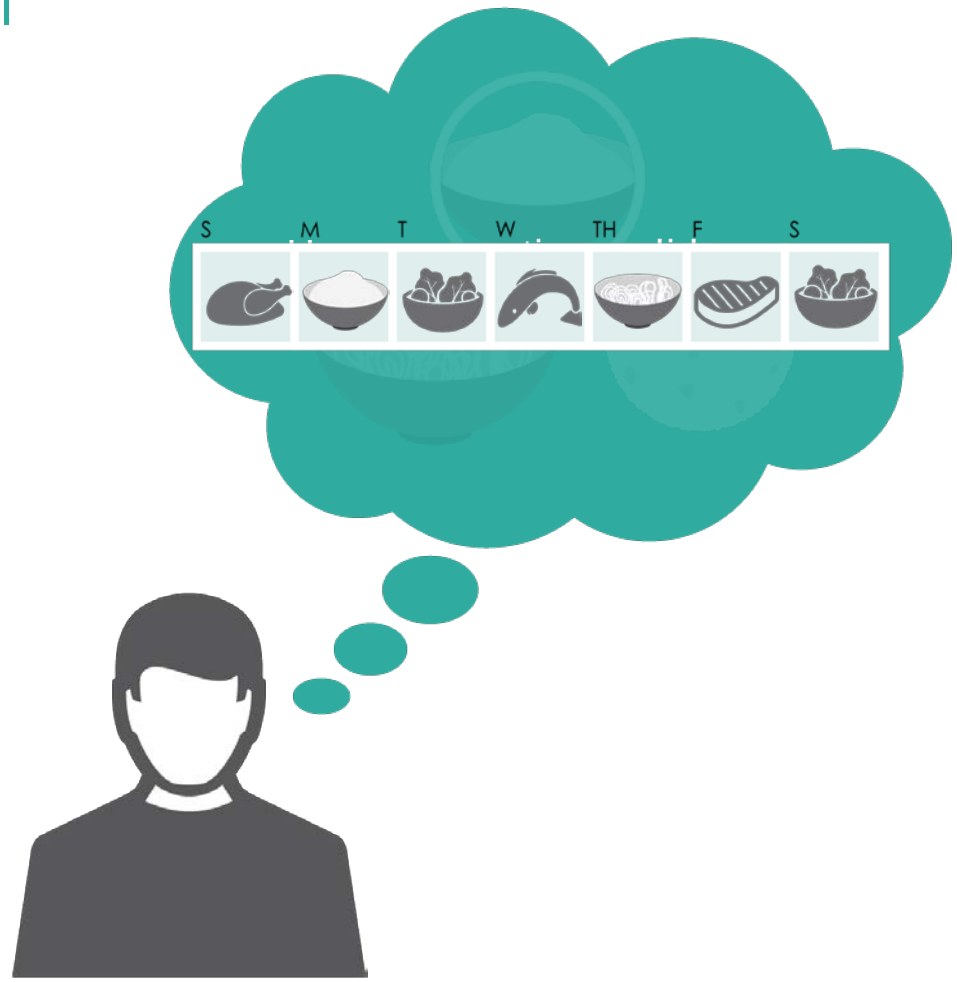
A. Prefer sitting

B. I don't prefer sitting



1. Comprehension of the question → **2. Retrieval of Information** → 3. Judgement and Estimation → 4. Reporting an Answer

Step 2: Retrieval



When you received your first measles vaccination, on a scale of 1-5, with 1 being painless, and 5 being unbearable painful: what was the level of pain?

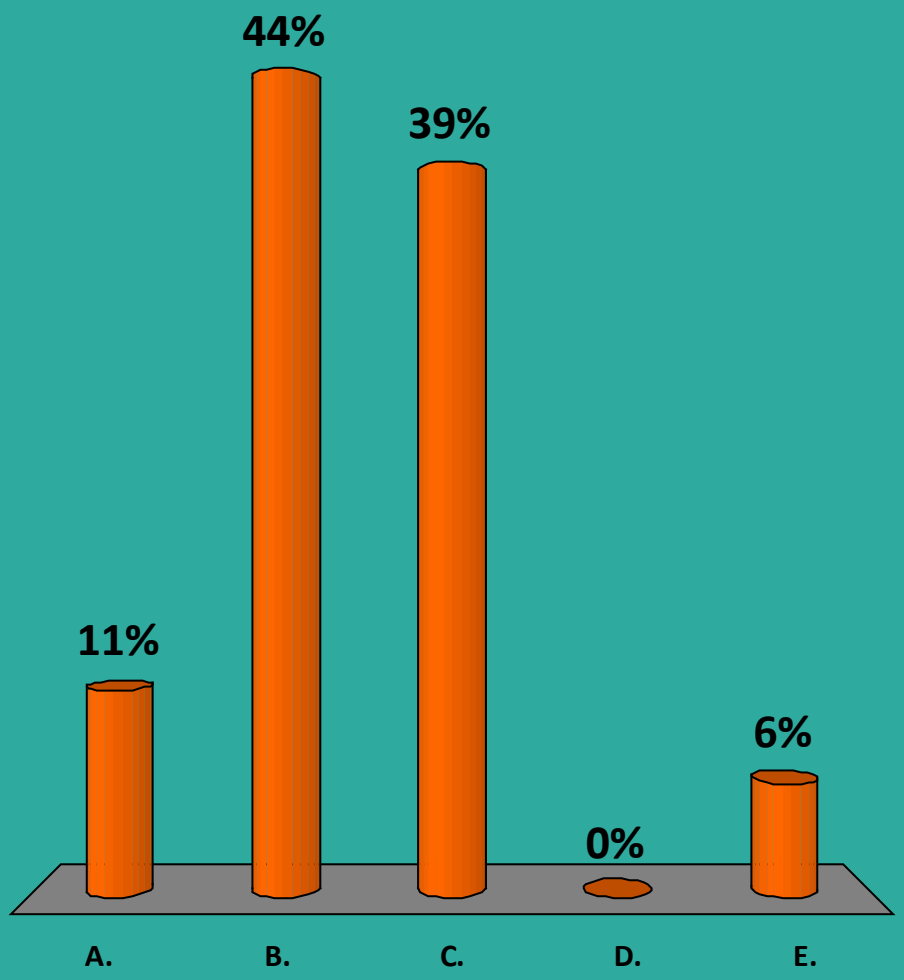
A. 1

B. 2

C. 3

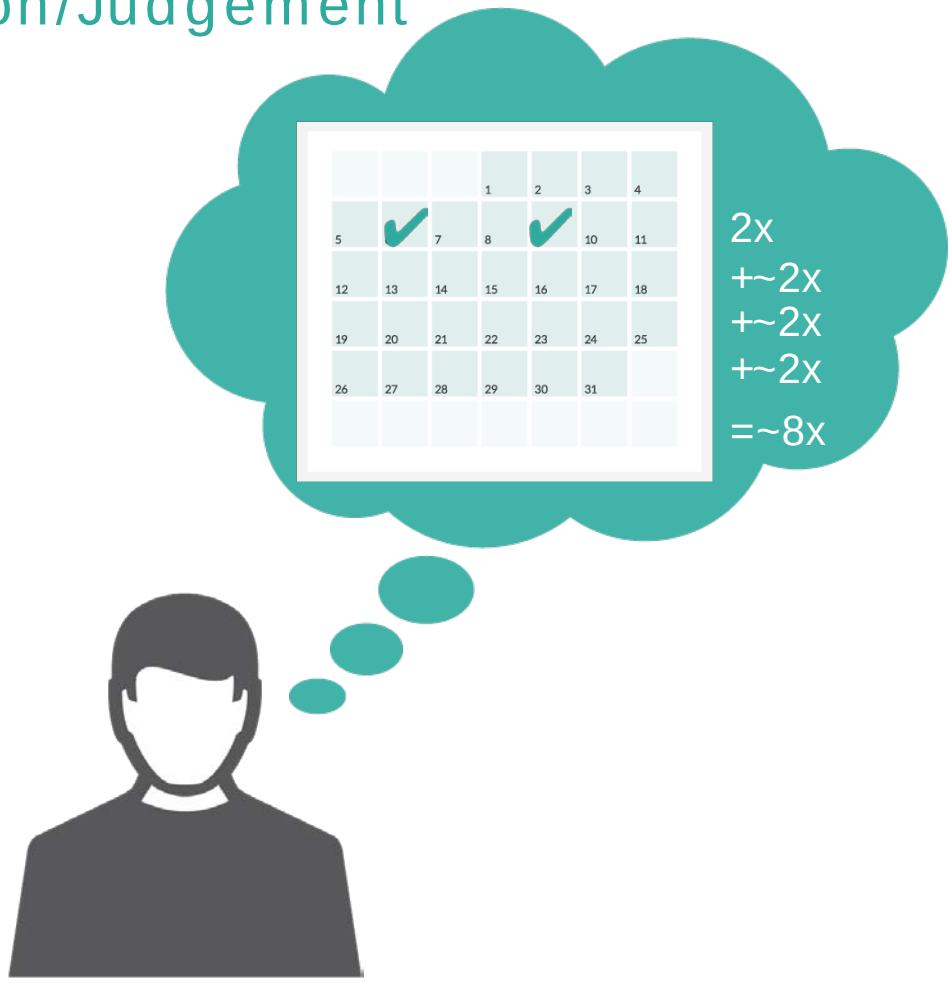
D. 4

E. 5



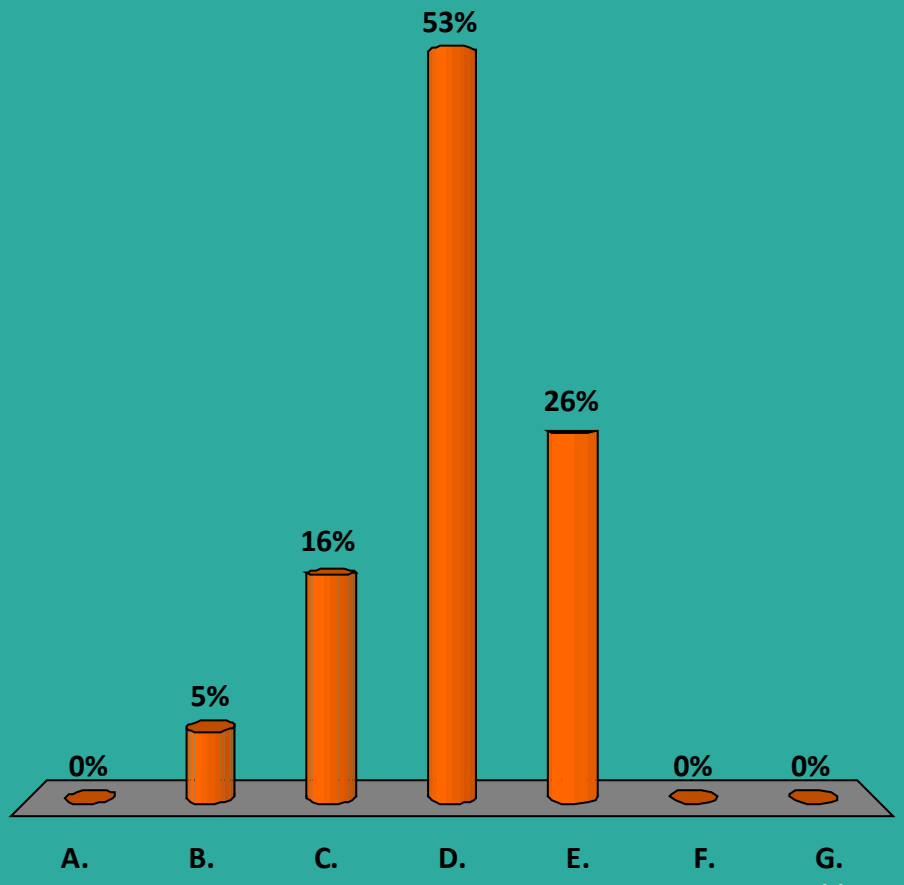
1. Comprehension of the question → 2. Retrieval of Information → **3. Judgement and Estimation** → 4. Reporting an Answer

Step 3: Estimation/Judgement



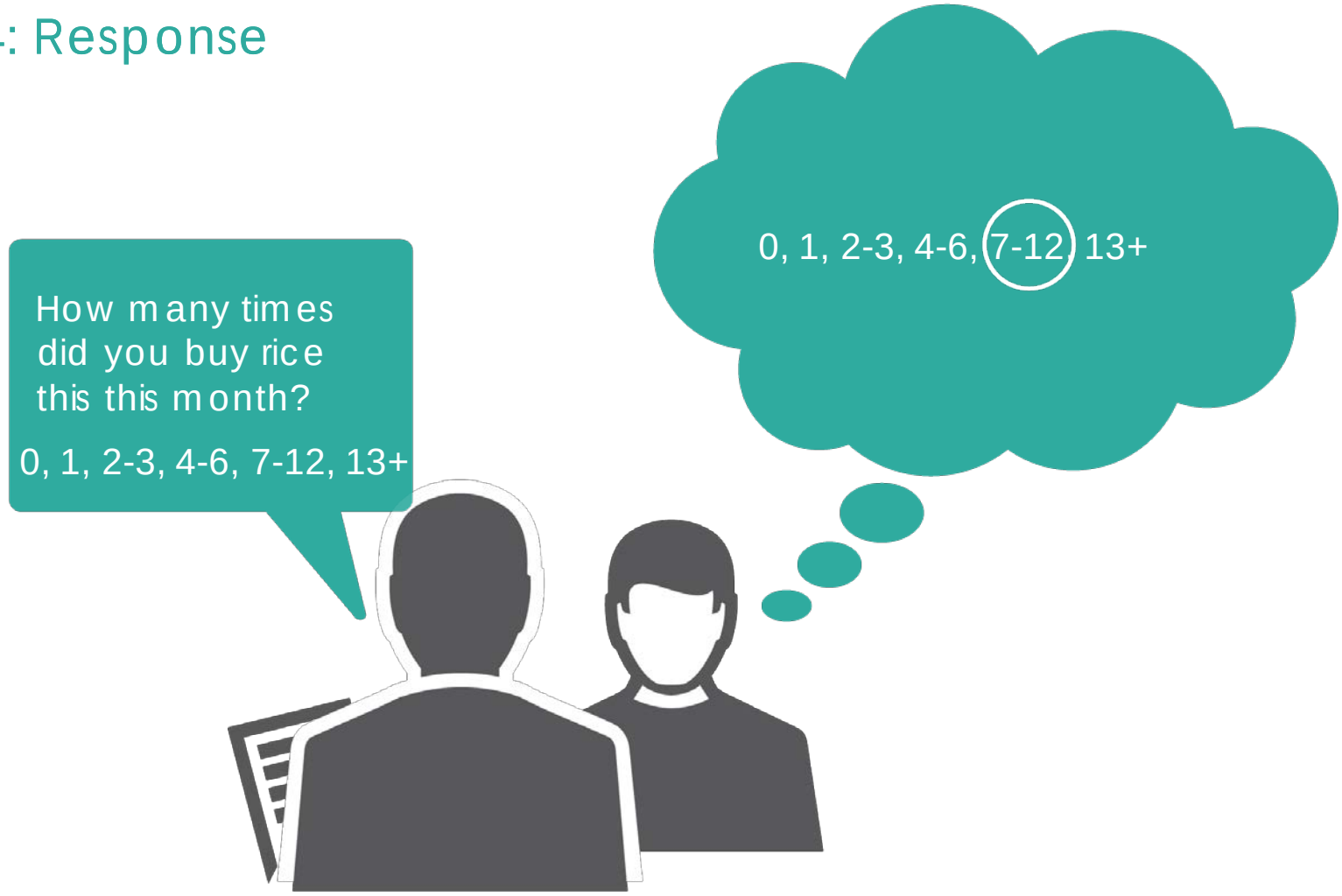
About how many calories do you think you consumed in your last large meal yesterday?

A. 0-99
B. 100-199
C. 200-499
D. 500-999
E. 1000-1499
F. 1500-2000
G. >2000



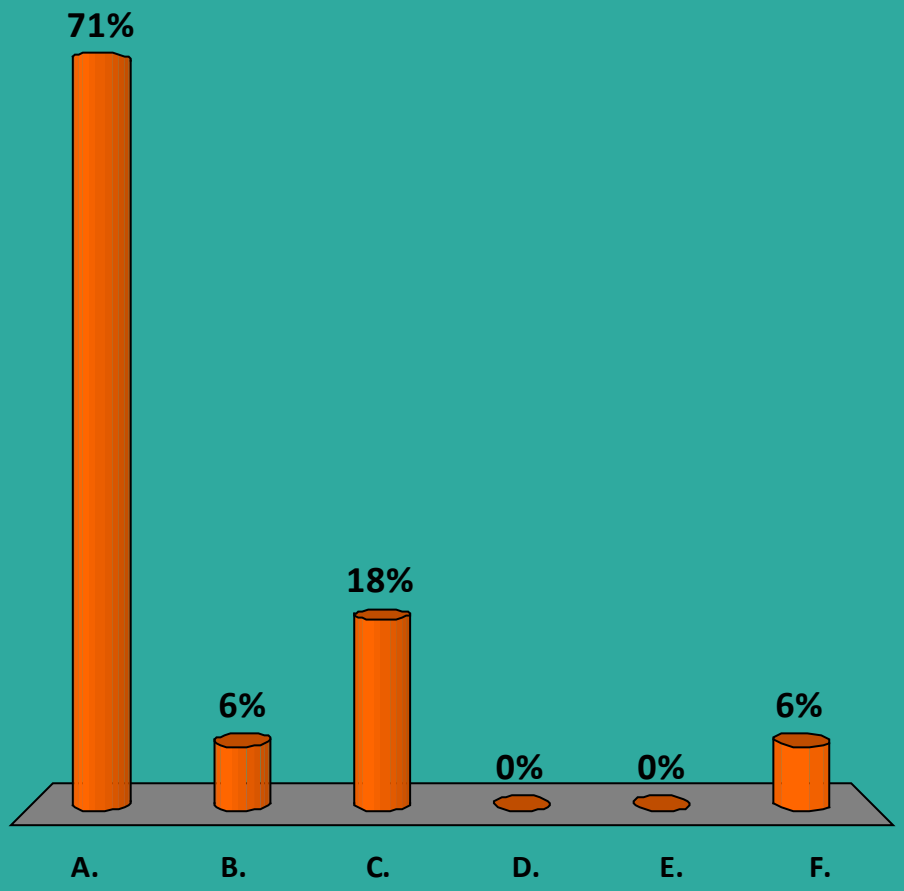
1. Comprehension of the question → 2. Retrieval of Information → 3. Judgement and Estimation → **4. Reporting an Answer**

Step 4: Response



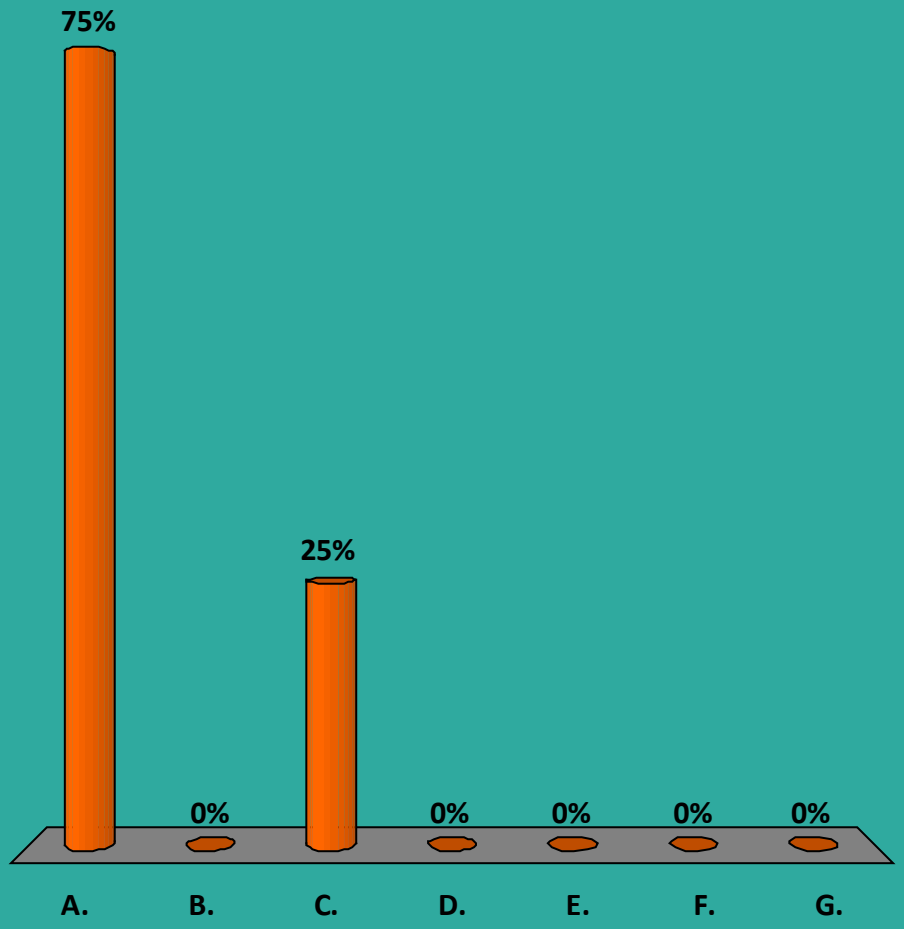
How many days have you taken illegal drugs in the past 12 months?

A. Never
B. Tried once: 1 time
C. Tried twice: 2 times
D. Frequently: 3 times
E. I am a drug user: 4 times
F. I am a drug addict: >4 times



How many days have you taken illegal drugs in your life?

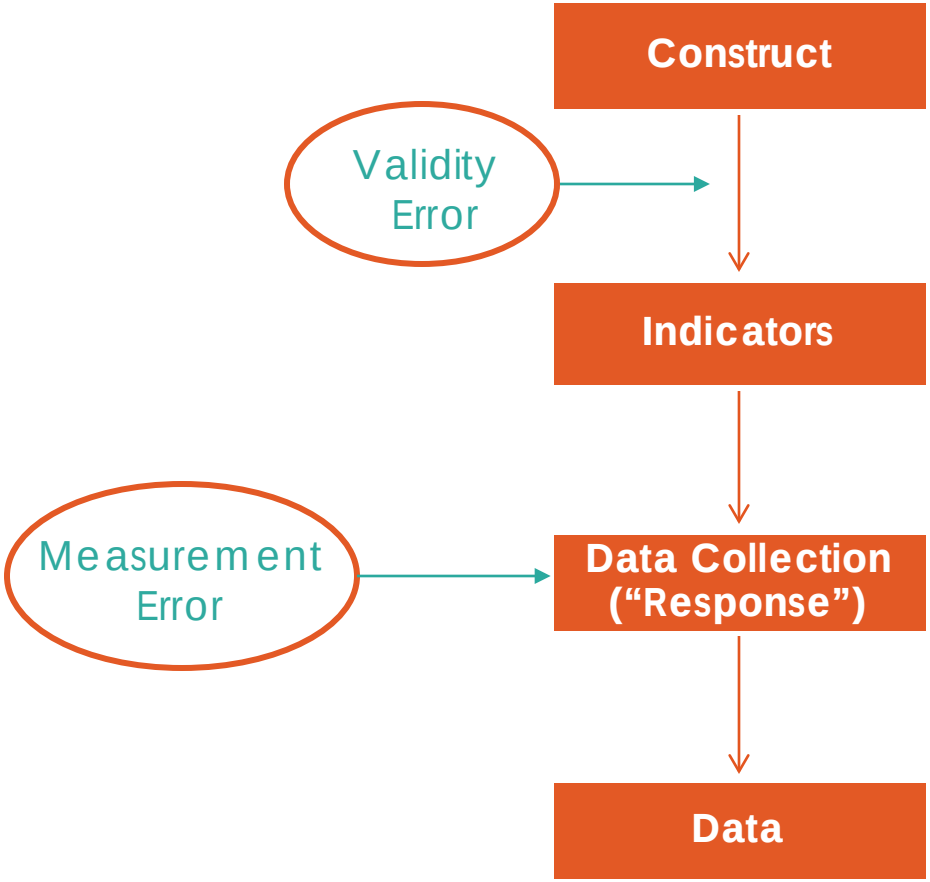
A. 0
B. 1-100
C. 101-1000
D. 1001-10,000
E. 10,001-20,000
F. 20,001-30,000
G. >30,000



How to Measure

Measurement Error

Error in Measurement



Where could the following question first produce error?

A. Validity

B. Comprehension

C. Retrieval

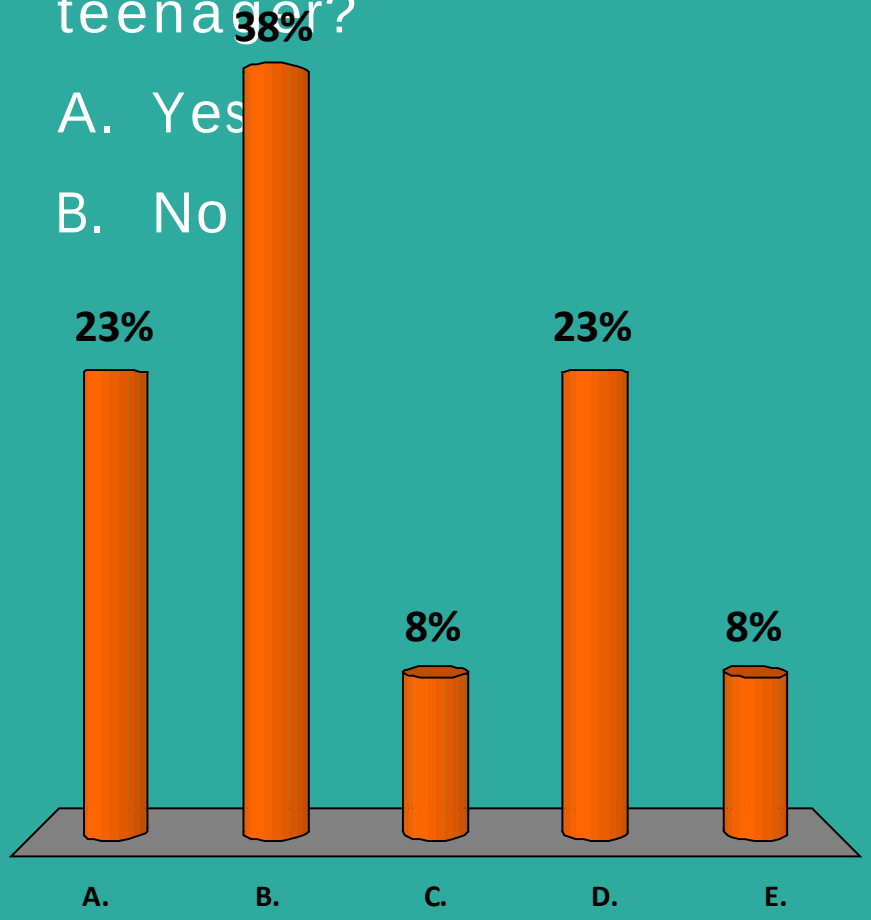
D. Judgment/ Estimation

E. Response

Q. Do you live with a teenager?

A. Yes

B. No



Measurement Error: Vagueness

Vague concepts where respondents may interpret the question in different ways.

Example:

Q. Do you live with a teenager?

- Yes
- No

Between what age ranges is a teenager?

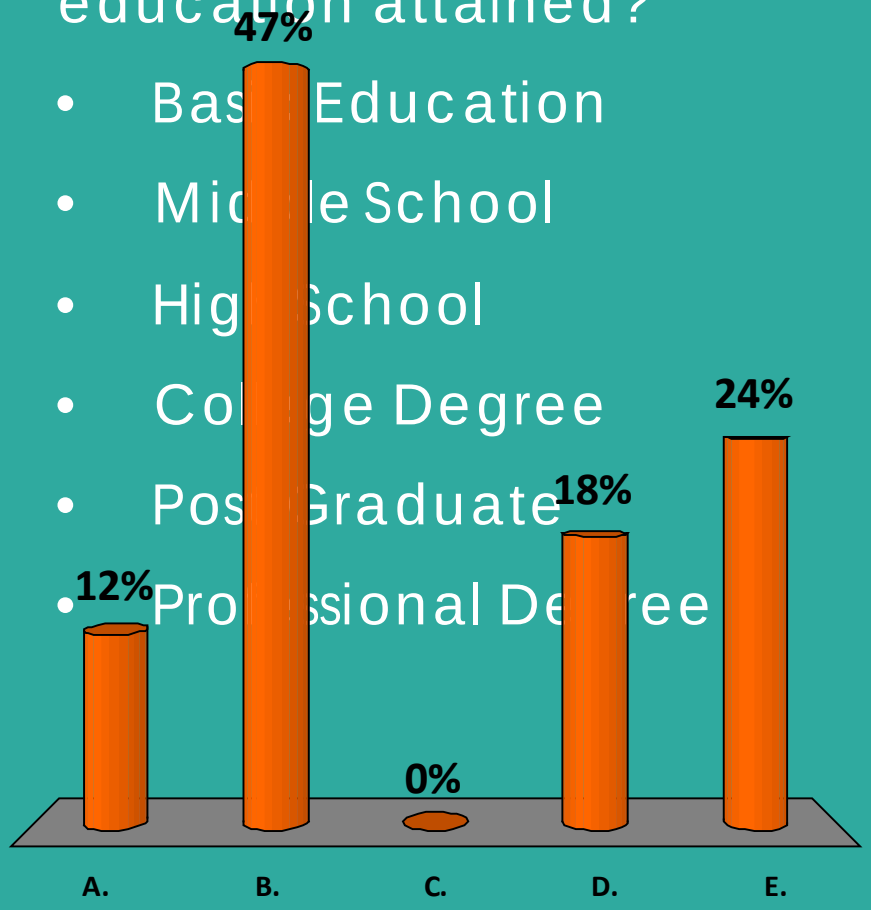
Make sure to define vague concepts

Where could the following question produce error?

A. Validity
B. Comprehension
C. Retrieval
D. Judgment/ Estimation
E. Response

Q. What is the level of education attained?

- Bas Education
- Mid le School
- Hig School
- Co ge Degree
- Pos Graduate
- Pro ssional De ree



Measurement Error: Completeness

The response categories do not include all categories that can be expected as a response

Example:

Q. What is the highest level of education completed?

- Basic Education (1-5th)
- Middle School (6th-8th)
- High School (9th-12th)
- College Degree
- Post Graduate
- Other Professional Degree (e.g. Medical, Law, Teacher)

"No education" or "vocational degree" is not a response

Pilot question to make sure that categories are exhaustive

Where could the following question first produce error?

A. Validity

B. Comprehension

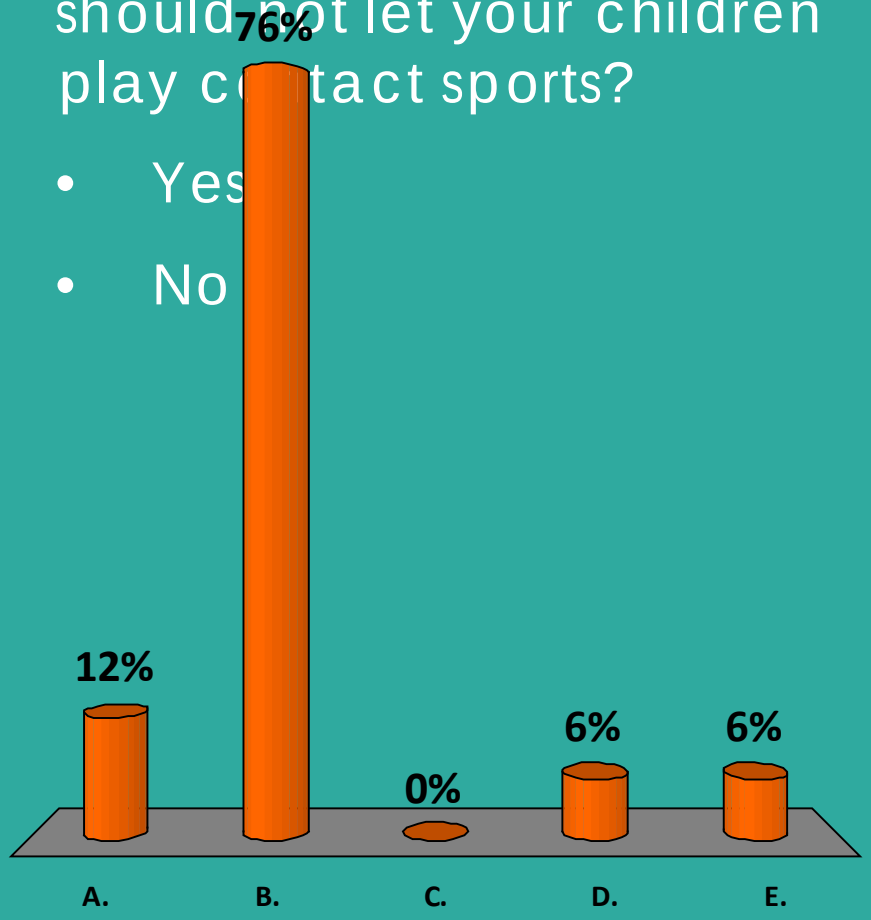
C. Retrieval

D. Judgment/ Estimation

E. Response

Q. Do you think that you should not let your children play contact sports?

- Yes
- No



Measurement Error: Negatives

Questions that include negatives can be confusing to the respondent and lead to misinterpretations.

Example:

Q. Do you think that you should not let your children play contact sports?

- Yes
- No

Having a negative might throw some people off

Avoid unnecessary negatives

Where could the following question first produce error?

A. Validity

B. Comprehension

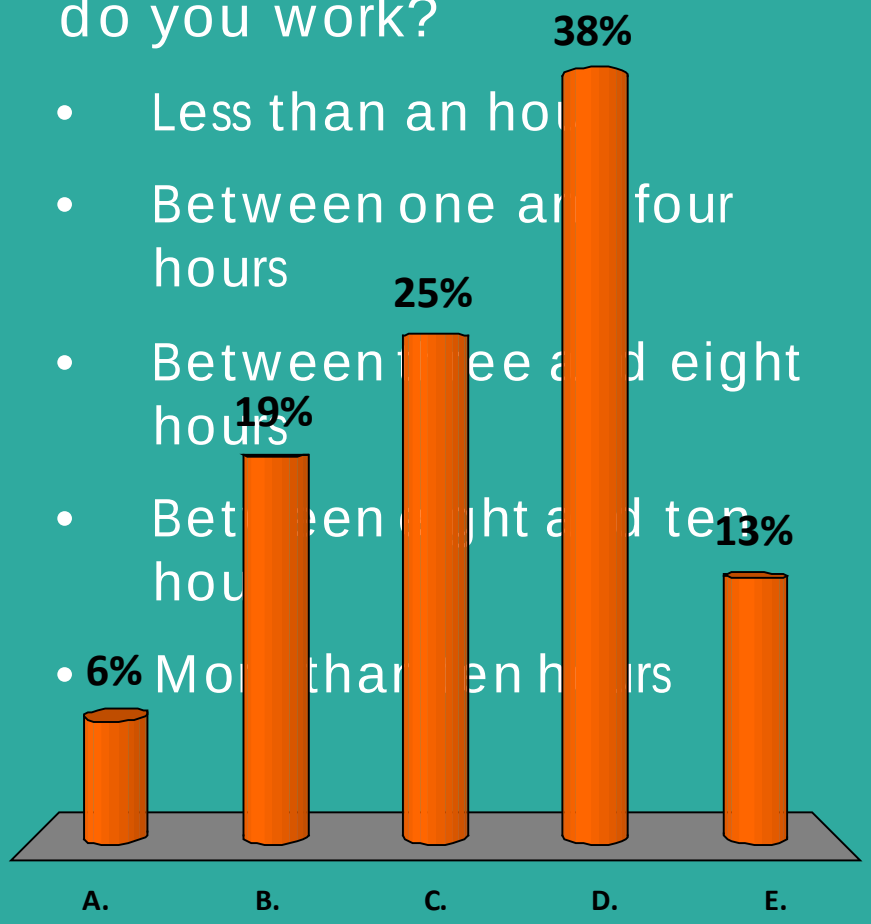
C. Retrieval

D. Judgment/ Estimation

E. Response

Q. How many hours a day do you work?

- Less than an hou
- Between one ar four hours
- Between ee a d eight hours
- Bet een ght a d ten hou
- Mo than en h rs



Measurement Error: Overlapping Categories

The categories overlap each other.

Example:

Q. How many hours a day do you work?

- Less than an hour
- Between one and four hours
- Between three and eight hours
- Between eight and ten hours
- More than ten hours

What would a person who works eight hours a day reply?

Make sure that all categories are mutually exclusive

Where could the following question first produce error?

A. Validity

B. Comprehension

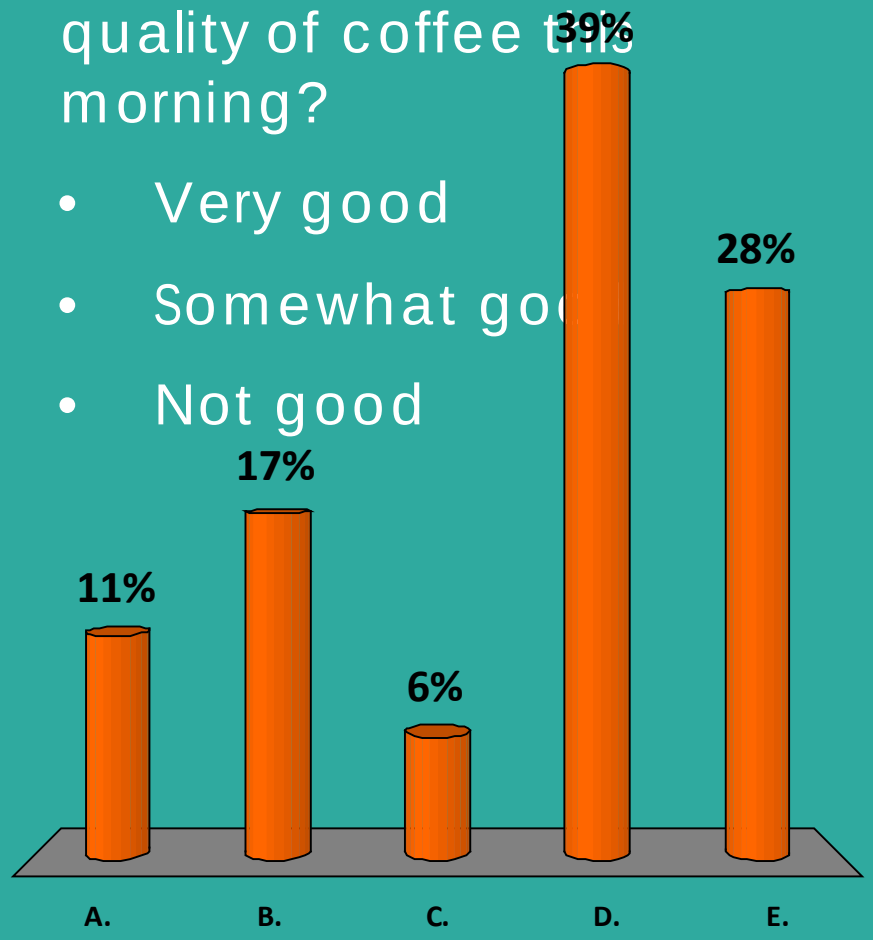
C. Retrieval

D. Judgment/ Estimation

E. Response

Q. How would you rate the quality of coffee this morning?

- Very good
- Somewhat good
- Not good



Measurement Error: Presumptions

The question assumes certain things about the respondent

Example:

Q. How would you rate the quality of coffee this morning?

- Very good
- Somewhat good
- Not good

We are assuming that the respondent drank the coffee

Use filters and skip patterns

Where could the following question first produce error?

A. Validity

B. Comprehension

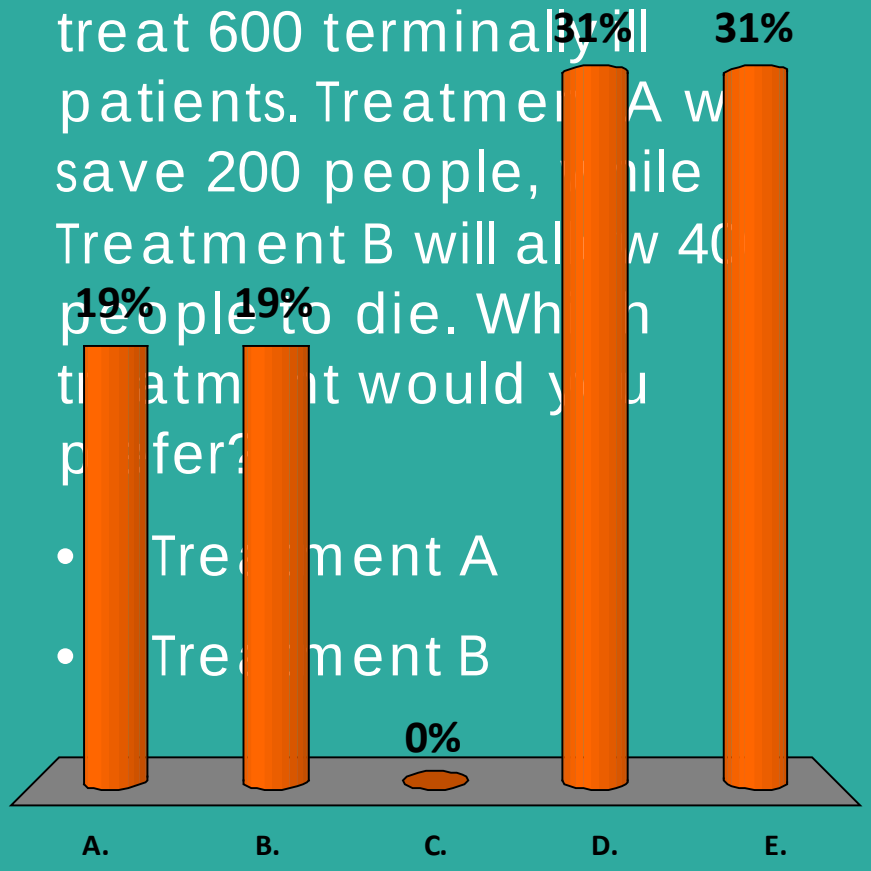
C. Retrieval

D. Judgment/ Estimation

E. Response

Q. Two new treatments have been developed to treat 600 terminally ill patients. Treatment A will save 200 people, while Treatment B will allow 400 people to die. Which treatment would you prefer?

- Treatment A
- Treatment B



Measurement Error: Framing effect

People react to a particular choice in different ways depending on how it is presented i.e. prefer gains over losses

Example:

Q. Two new treatments have been developed to treat 600 terminally ill patients. Treatment A will save 200 people, while Treatment B will allow 400 people to die. Which treatment would you prefer?

- Treatment A
- Treatment B

Treatment A is preferable because it has been framed as a gain

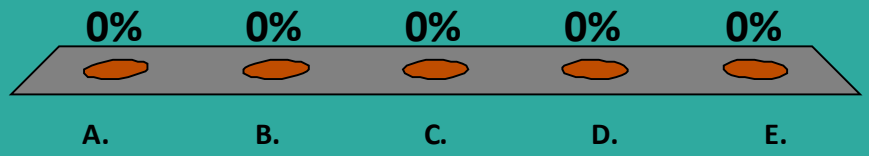
Try to be neutral when framing questions

Where could the following question first produce error?

A. Validity
B. Comprehension
C. Retrieval
D. Judgment/ Estimation
E. Response

Q. How long did you have to wait last time you voted?

- No time (there was no line, or I voted by mail)
- Less than 10 minutes
- Between 10 minutes and 30
- More than 30 minutes but less than an hour
- An hour or more



Measurement Error: Recall Bias

People may retrieve recollections regarding events or experiences differently

Example:

Q. How long did you have to wait last time you voted?

- No time (there was no line, or I voted by mail)
- Less than 10 minutes
- Between 10 minutes and 30
- More than 30 minutes but less than an hour
- An hour or more

This experience may be more vivid for some respondents than others.

You can ask respondents to keep a diary or save their receipts

Where could the following question first produce error?

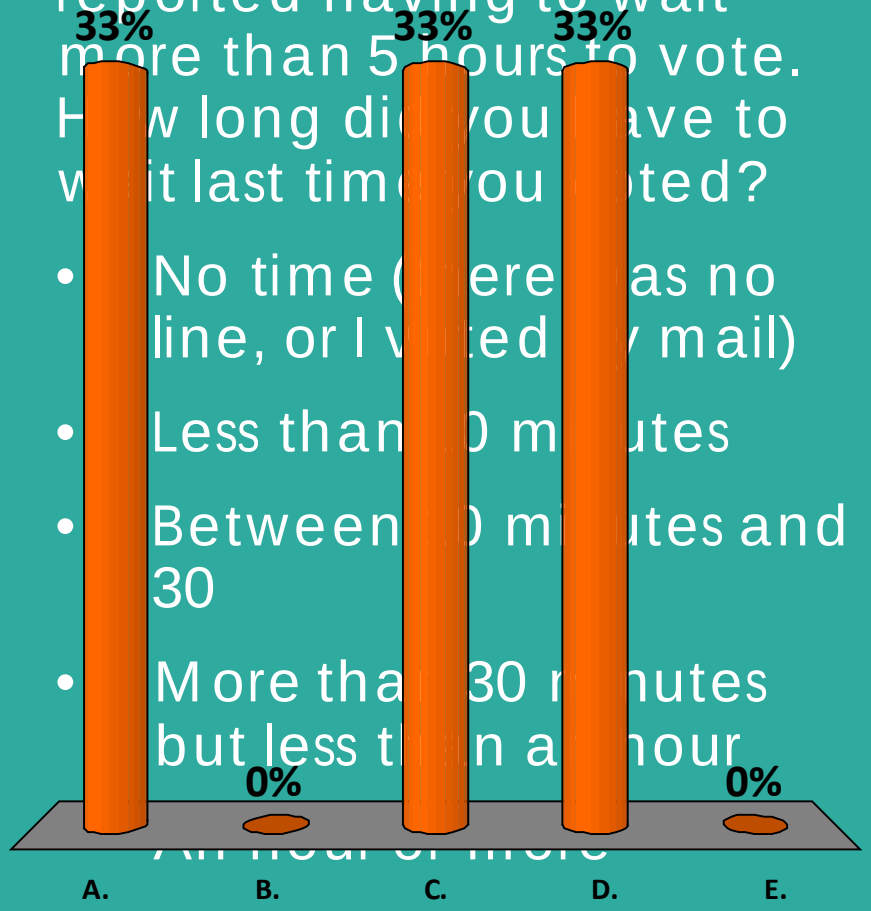
A. Validity

B. Comprehension

C. Retrieval

D. Judgment/ Estimation

E. Response



Measurement Error: Anchoring Bias

People tend to rely too heavily on the first piece of information seen

Example:

Q. In Arizona, some voters reported having to wait more than 5 hours to vote. How long did you have to wait last time you voted?

- No time (there was no line, or I voted by mail)
- Less than 10 minutes
- Between 10 minutes and 30
- More than 30 minutes but less than an hour
- An hour or more

Respondents will be more likely to give a number on the higher end of the spectrum

Avoid adding anchors to your questions

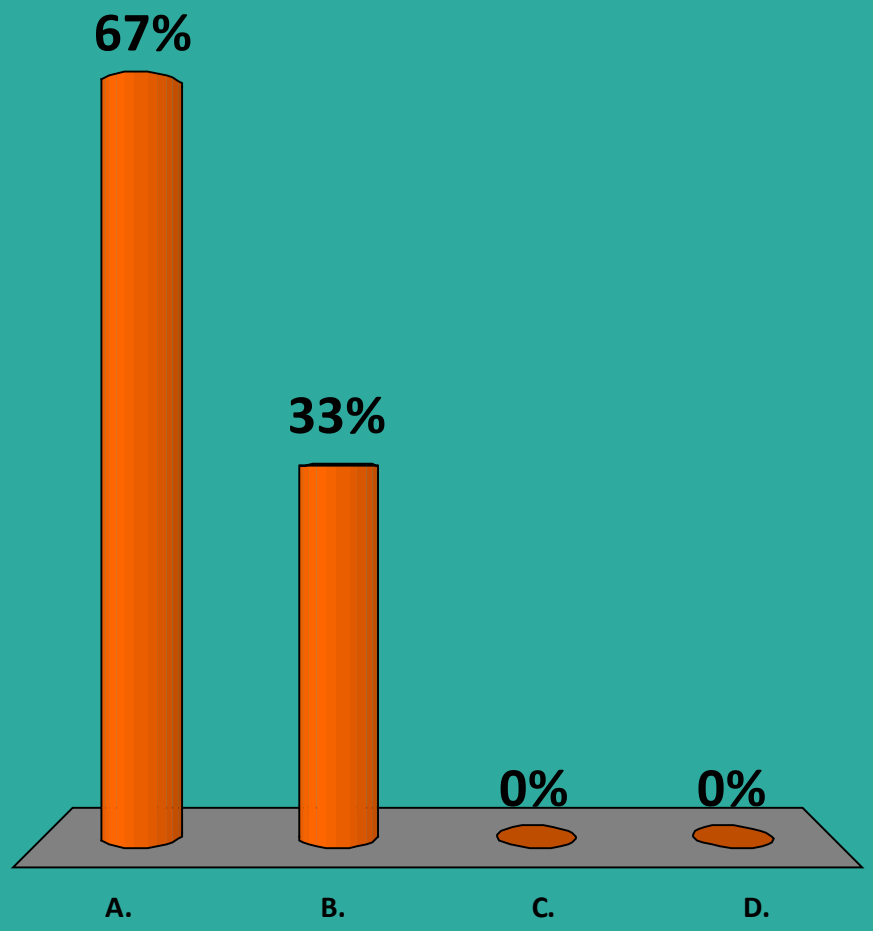
How many meals have you eaten in the past one hour?

A. 0

B. 1

C. 2

D. 3



Measurement Error: Telescoping Bias

People perceive recent events as being more remote than they are (backward telescoping) and distant events as being more recent than they are (forward telescoping)

Example:

Q. Did you purchase a TV or other electronic (worth over $500) in the past 12 months?

_____ emails

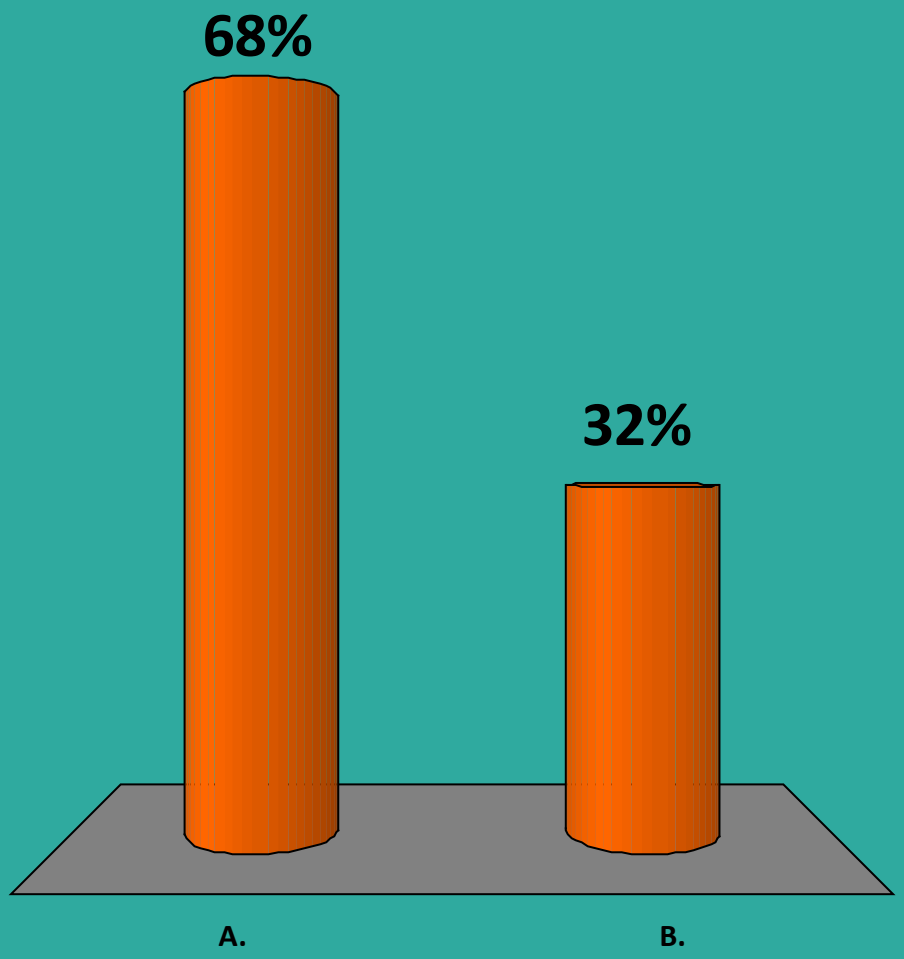
This will lead to over reporting due to forward telescoping of events that happened before 12 months ago

Visit once at the beginning of the reference period. Then ask, "since the last time I visited you, have you…?"

In the past year, have you said anything disparaging about academics to your colleagues?

A. Yes

B. No



Measurement Error: Social Desirability Bias

Tendency of respondents to answer questions in a manner that is favorable to others i.e. emphasize strengths, hide flaws, or avoid stigma

Example:

Q. Do you beat your wife?

- Yes
- No

Respondents would be shy to admit to such behavior

Ask indirectly, ensure privacy

Sources Measurement Error

- Completeness
- Vagueness
- Negatives
- Overlapping Categories
- Presumptions
- Framing effect
- Recall bias
- Anchoring bias
- Telescoping bias
- Social desirability bias

How to Measure

Best Practices

Tips for designing questions

- Break complex questions into smaller questions, asking only one question at a time.
- With closed questions, include all reasonable possibilities as explicit response options.
- Make questions as specific as possible (not for sensitive questions).
- Use long instead of short questions (for sensitive questions).

Tips for designing questions

- Use familiar words to describe sensitive behaviors.
- Include the sensitive question with other sensitive questions so that it stands out less.
- Use visual cues to convey certain concepts (social cohesion, pain indicator, happiness).
- Use visuals consistently to define the desired path through the questionnaire. (self-administered questions)

Tips for designing questions

- Use words that all respondents can understand.
- The first questions should be easy and pleasant to answer and should serve to build trust between the interviewee and the researcher.
- The questions on the same topic should be grouped.
- Should include filters, to avoid asking respondents questions that do not apply to them.

Tips for piloting your survey

- Pretest: procedures to determine whether the questionnaire works.
  - It is a small group survey representative of the target population after a group interview is done.

- Cognitive Interview: looking to find how respondents understand the questions. After asking the question, ask them feedback on their question, probing all 4 steps of the response process.

- Expert Review: Ask measurement experts. But also compare to well-established, well-vetted, tried and tested surveys

Tips for piloting your survey

- No answer: if a certain question has a high number of omissions indicates that something is wrong.
- Multiple answers: questions where there is a single answer respondents placed more than one.
- Answer "other": high response rate in this category indicates that the answers offered are not exhaustive.

Other things to consider

- Question wording, definitions, recall period
- Answer choice
  - Open/closed, single v. multiple options, units, likert/scale, index, visual cues
- Translation
  - Back-translate and pretest in local languages
- Surveyor training/quality
- Data entry
- Length, fatigue

Thank You!