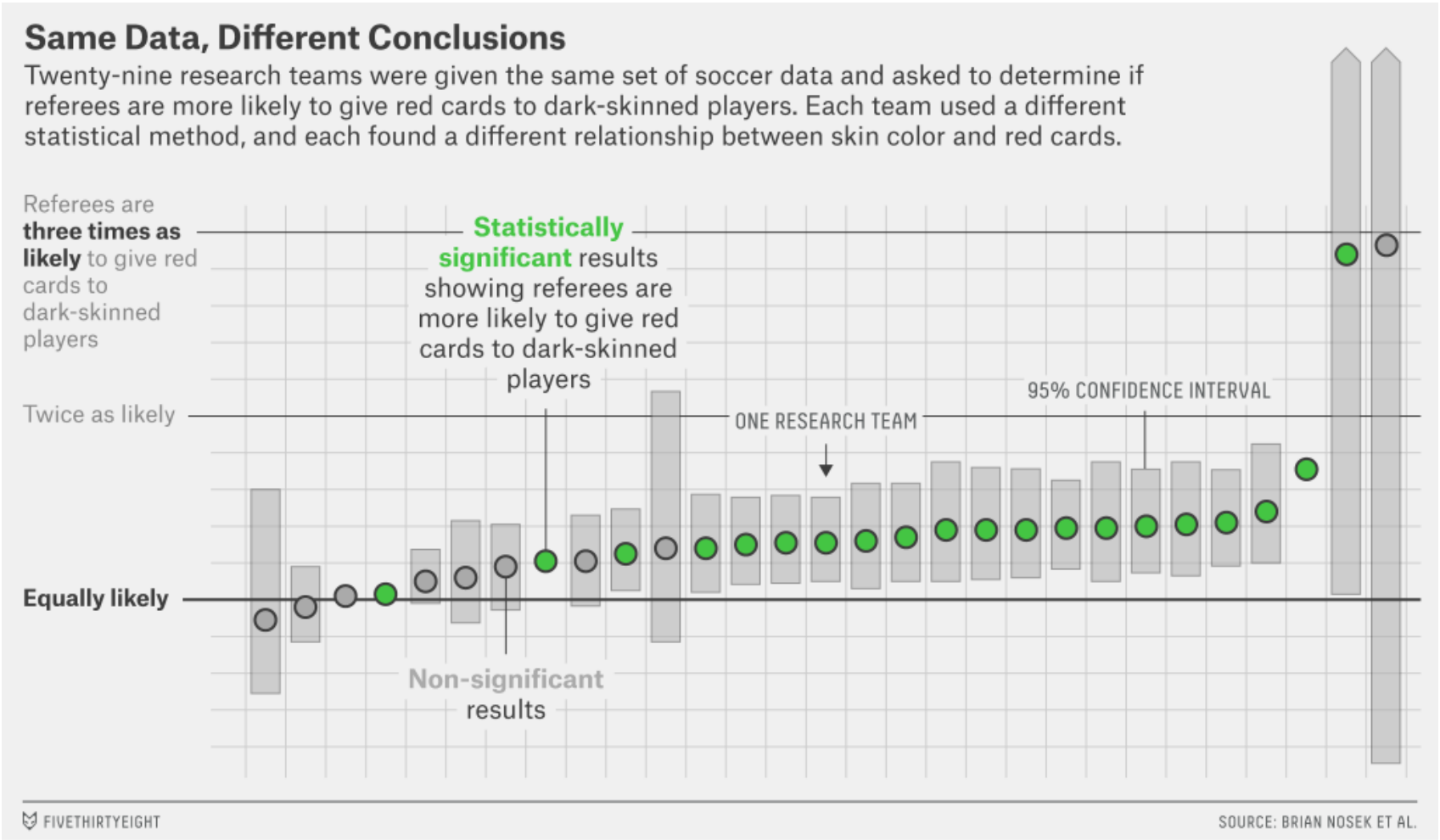
Researcher Choices

Data Analysis for Journalism and Political Communication
(Spring 2026)

Prof. Bell


Same Data, Different Conclusions
Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.
Referees are three times as likely to give red cards to dark-skinned players
Twice as likely
Equally likely
Statistically significant results showing referees are more likely to give red cards to dark-skinned players
Non-significant results
ONE RESEARCH TEAM
95% CONFIDENCE INTERVAL
FIVETHIRTYEIGHT
SOURCE: BRIAN NOSEK ET AL.

Researcher Choices

1. What is my hypothesis?

Researcher Choices

1. What is my hypothesis?
2. What do I measure?

Researcher Choices

1. What is my hypothesis?
2. What do I measure?
3. How do I collect my data?

Researcher Choices

1. What is my hypothesis?
2. What do I measure?
3. How do I collect my data?
4. How much data do I collect?

Researcher Choices

1. What is my hypothesis?
2. What do I measure?
3. How do I collect my data?
4. How much data do I collect?
5. What statistical analyses do I use?

Researcher Choices

1. What is my hypothesis?
2. What do I measure?
3. How do I collect my data?
4. How much data do I collect?
5. What statistical analyses do I use?
6. How do I handle outliers, missing data, and other peculiarities?

What is my hypothesis?

Choosing a hypothesis is all about avoiding error:

- **Type I error**: False positives
  - Sending the innocent to jail - this is bad!

What is my hypothesis?

Choosing a hypothesis is all about avoiding error:

- **Type I error**: False positives
  - Sending the innocent to jail - this is bad!
- **Type II error**: False negatives
  - Letting the guilty go free - we can accept this

What is my hypothesis?

Choosing a hypothesis is all about avoiding error:

- **Type I error**: False positives
  - Sending the innocent to jail - this is bad!
- **Type II error**: False negatives
  - Letting the guilty go free - we can accept this

Our goal is to reduce Type I error. Assume that the data is innocent (that the hypothesis is false) until it is proven guilty.

Measuring Type I Error

P-value

The chance that we would get a particular result from our test if the true answer is false

Measuring Type I Error

P-value

The chance that we would get a particular result from our test if the true answer is false

- The p-value is our chance of committing a Type I error - sending the innocent to jail

Measuring Type I Error

P-value

The chance that we would get a particular result from our test if the true answer is false

- The p-value is our chance of committing a Type I error - sending the innocent to jail
- Common p-value cut-offs in scientific research: .01, .05, and .1 indicate **statistical significance**

Measuring Type I Error

Hypothesis: A student cheated on an exam.

P-value (chance that we conclude that student cheated, but they did not): .50

Measuring Type I Error

Hypothesis: A student cheated on an exam.

- The student performed much better on this exam than on previous exams

P-value (chance that we conclude that student cheated, but they did not): .25

Measuring Type I Error

Hypothesis: A student cheated on an exam.

- The student performed much better on this exam than on previous exams
- The student finished their exam more quickly than other students

P-value (chance that we conclude that student cheated, but they did not): .15

Measuring Type I Error

Hypothesis: A student cheated on an exam.

- The student performed much better on this exam than on previous exams
- The student finished their exam more quickly than other students
- The student's roommate saw them up all night studying before the exam

P-value (chance that we conclude that student cheated, but they did not): .40

Measuring Type I Error

Hypothesis: A student cheated on an exam.

- The student performed much better on this exam than on previous exams
- The student finished their exam more quickly than other students
- The student's roommate saw them up all night studying before the exam
- The student missed the same questions as other students

P-value (chance that we conclude that student cheated, but they did not): .75

The Scientific Method

- P-values are a product of the data we use and our choices about what we include and exclude from the analysis.

The Scientific Method

- P-values are a product of the data we use and our choices about what we include and exclude from the analysis.
- We follow the scientific method: Theory $\Rightarrow$ Hypothesis $\Rightarrow$ Test $\Rightarrow$ Analyze $\Rightarrow$ Report

The Scientific Method

- P-values are a product of the data we use and our choices about what we include and exclude from the analysis.
- We follow the scientific method: Theory $\Rightarrow$ Hypothesis $\Rightarrow$ Test $\Rightarrow$ Analyze $\Rightarrow$ Report
- But in practice, no analysis plan survives contact with the data

Are Democrats or Republicans Good for the Economy?

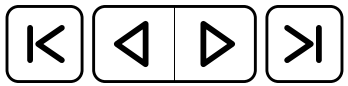
Use (a recreation of) FiveThirtyEight's online tool to test what you think is the best approach to answering the question. There are no right or wrong answers - just select the model you think is best, and report your results in the form:

https://bit.ly/smpa2152

(The link to the tool is on the form.)

Are Democrats or Republicans Good for the Economy?

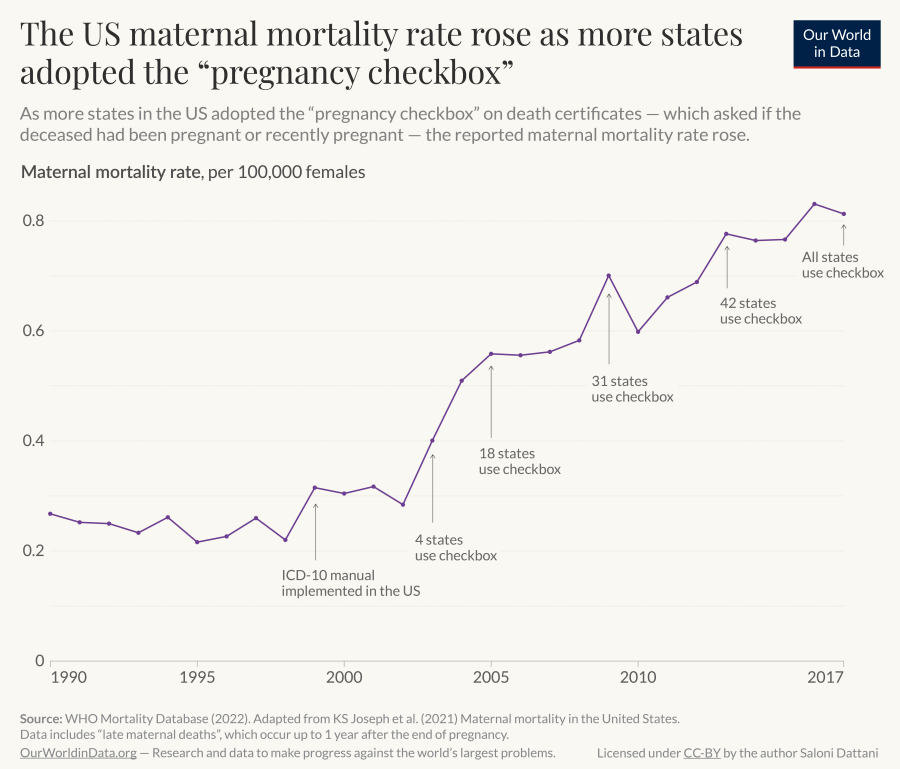
Controls:

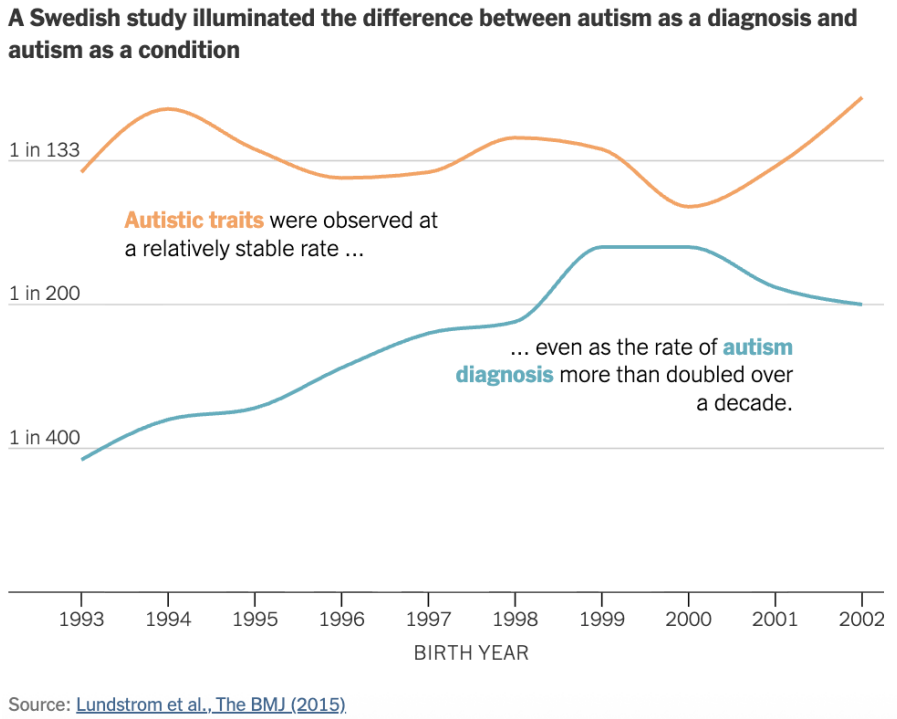


What do I measure?

Operationalization

The process of defining a measurable version of a concept.





Principles of Good Operationalization

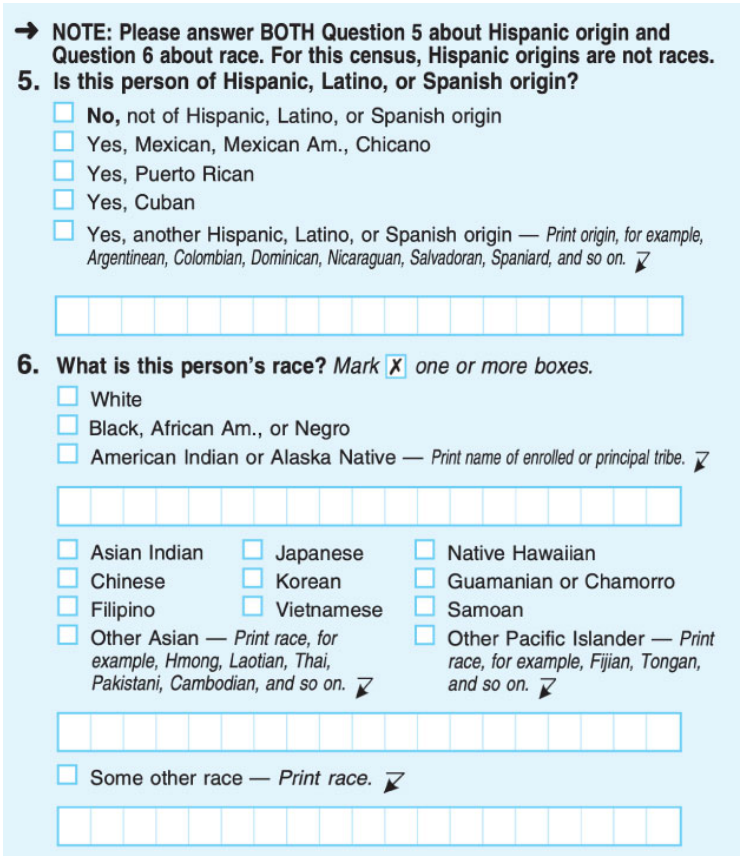
1. Unambiguous

Principles of Good Operationalization

1. Unambiguous
2. Parsimonious

Principles of Good Operationalization

1. Unambiguous
2. Parsimonious
3. Accurate



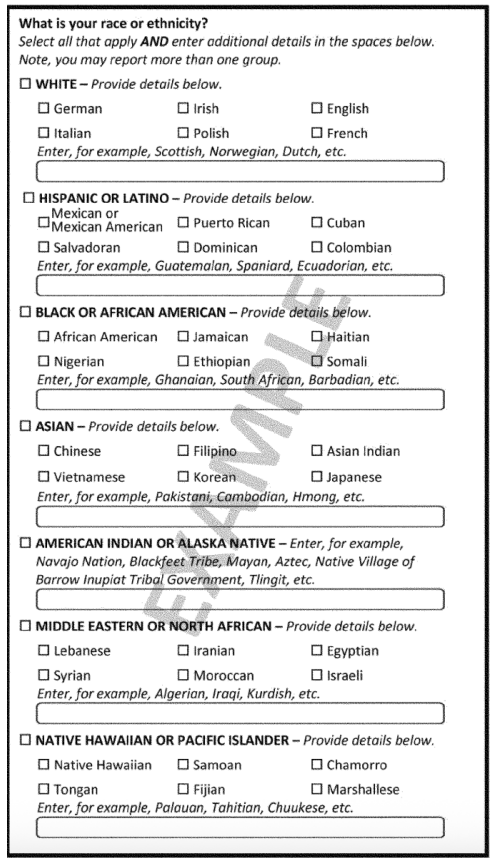
➜ **NOTE: Please answer BOTH Question 5 about Hispanic origin and Question 6 about race. For this census, Hispanic origins are not races.**

5. Is this person of Hispanic, Latino, or Spanish origin?

- ☐ **No,** not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

6. What is this person's race? *Mark* ☒ *one or more boxes.*

- ☐ White
- ☐ Black, African Am., or Negro
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
- ☐ Asian Indian
- ☐ Chinese
- ☐ Filipino
- ☐ Japanese
- ☐ Korean
- ☐ Vietnamese
- ☐ Native Hawaiian
- ☐ Guamanian or Chamorro
- ☐ Samoan
- ☐ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
- ☐ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*
- ☐ Some other race — *Print race.*



What is your race or ethnicity?
Select all that apply ***AND*** *enter additional details in the spaces below.*
Note, you may report more than one group.

☐ **WHITE** – *Provide details below.*
☐ German ☐ Irish ☐ English
☐ Italian ☐ Polish ☐ French
Enter, for example, Scottish, Norwegian, Dutch, etc.

☐ **HISPANIC OR LATINO** – *Provide details below.*
☐ Mexican or Mexican American ☐ Puerto Rican ☐ Cuban
☐ Salvadoran ☐ Dominican ☐ Colombian
Enter, for example, Guatemalan, Spaniard, Ecuadorian, etc.

☐ **BLACK OR AFRICAN AMERICAN** – *Provide details below.*
☐ African American ☐ Jamaican ☐ Haitian
☐ Nigerian ☐ Ethiopian ☐ Somali
Enter, for example, Ghanaian, South African, Barbadian, etc.

☐ **ASIAN** – *Provide details below.*
☐ Chinese ☐ Filipino ☐ Asian Indian
☐ Vietnamese ☐ Korean ☐ Japanese
Enter, for example, Pakistani, Cambodian, Hmong, etc.

☐ **AMERICAN INDIAN OR ALASKA NATIVE** – *Enter, for example, Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Tribal Government, Tlingit, etc.*

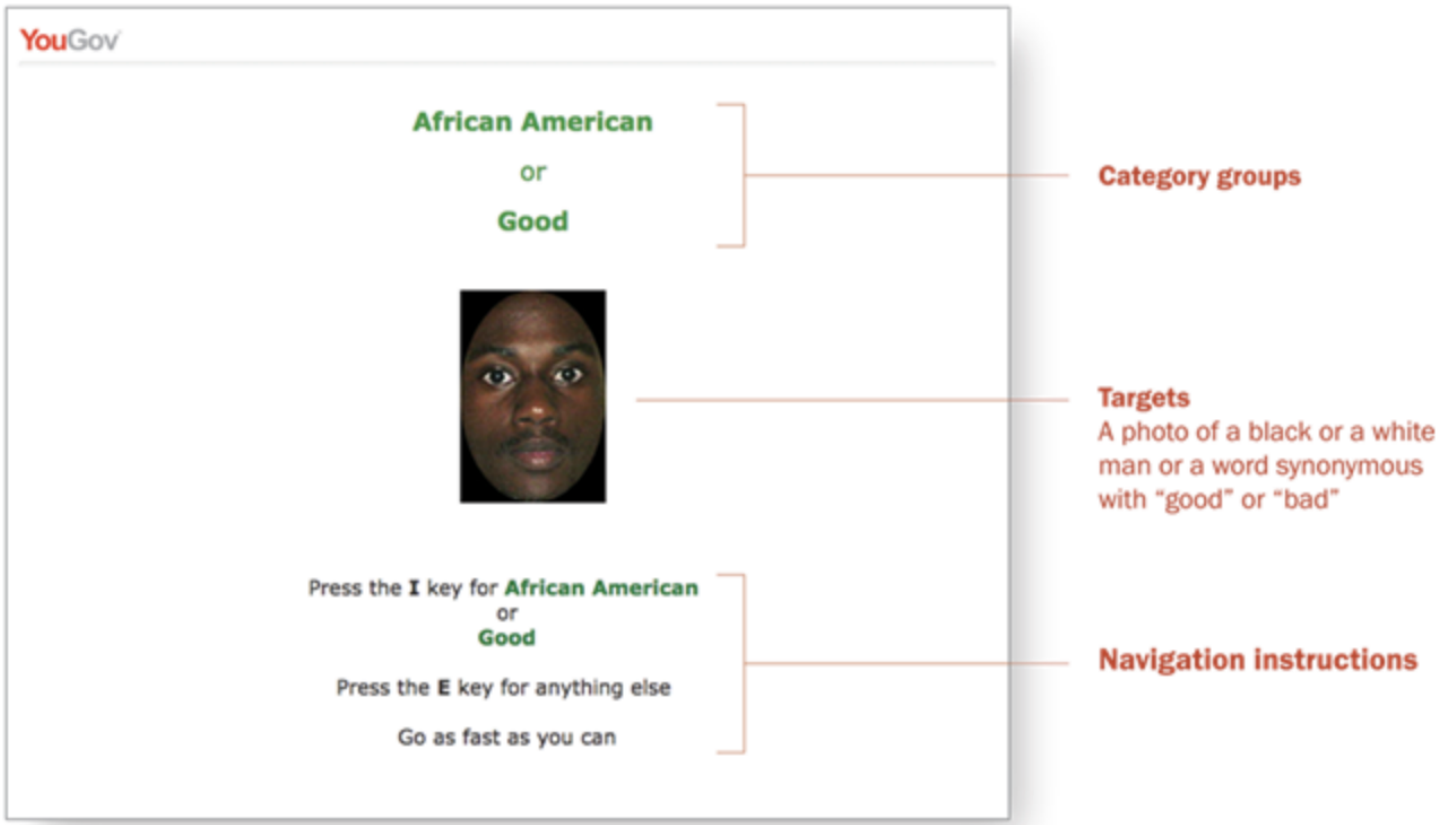
☐ **MIDDLE EASTERN OR NORTH AFRICAN** – *Provide details below.*
☐ Lebanese ☐ Iranian ☐ Egyptian
☐ Syrian ☐ Moroccan ☐ Israeli
Enter, for example, Algerian, Iraqi, Kurdish, etc.

☐ **NATIVE HAWAIIAN OR PACIFIC ISLANDER** – *Provide details below.*
☐ Native Hawaiian ☐ Samoan ☐ Chamorro
☐ Tongan ☐ Fijian ☐ Marshallese
Enter, for example, Palauan, Tahitian, Chuukese, etc.

EXAMPLE

Principles of Good Operationalization

1. Unambiguous
2. Parsimonious
3. Accurate
4. Reliable



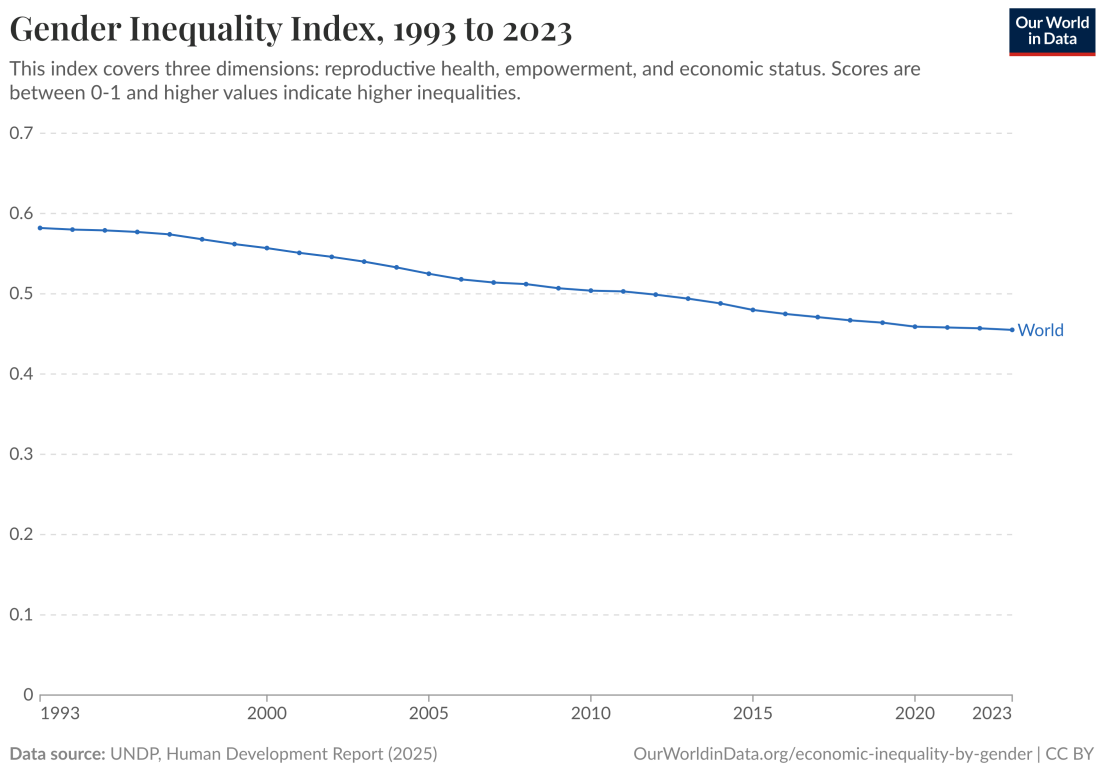
Principles of Good Operationalization

1. Unambiguous
2. Parsimonious
3. Accurate
4. Reliable
5. Feasible

Principles of Good Operationalization

Index (or Scale) Variable

An index (or scale) variable is a type of proxy measure in which the researcher combines multiple sources of data into a single representation of the concept under investigation using a pre-defined mathematical operation.



Exercise: Operationalization

1. You want to measure how happy people are
2. You want to measure people's driving ability
3. You want to measure the political ideology of a member of Congress

How do I collect my data?

Data Generating Process

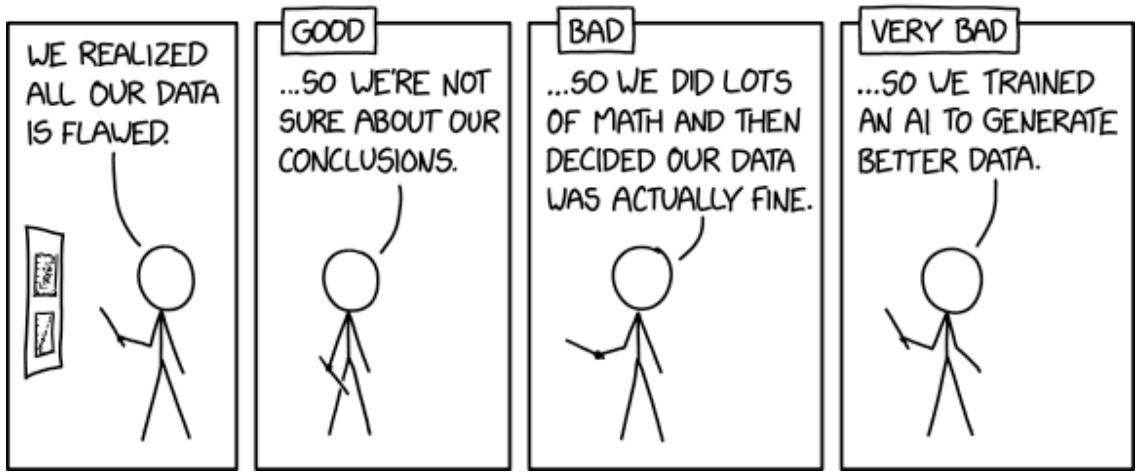
The rules and procedures that produce the data one is interested in

How do I collect my data?

Data Generating Process

The rules and procedures that produce the data one is interested in

- No amount of statistical wizardry can compensate for bad data



How do I collect my data?

Data Generating Process

The rules and procedures that produce the data one is interested in

- No amount of statistical wizardry can compensate for bad data
- The gold standard of data generating processes is the **random sample**

Sampling

- The group we are interested in studying is known as the **population**

Sampling

- The group we are interested in studying is known as the **population**
- Often, we are not able to count every unit in the population, so we take a **sample**

Sampling

- The group we are interested in studying is known as the **population**
- Often, we are not able to count every unit in the population, so we take a **sample**
- The key to any data analysis project is a quality sample, which is determined by two elements:

Sampling

- The group we are interested in studying is known as the **population**
- Often, we are not able to count every unit in the population, so we take a **sample**
- The key to any data analysis project is a quality sample, which is determined by two elements:
  1. A **random sample** of the population

Sampling

- The group we are interested in studying is known as the **population**
- Often, we are not able to count every unit in the population, so we take a **sample**
- The key to any data analysis project is a quality sample, which is determined by two elements:
  1. A **random sample** of the population
  2. The **sample size** is sufficiently large

Random Sample

Definition

The probability of any given unit being drawn from the population is uniform (the same)

Random Sample

Definition

The probability of any given unit being drawn from the population is uniform (the same)

- A failure of each unit to have a uniform probability of being drawn from the population is known as **selection bias**

Random Sample

Definition

The probability of any given unit being drawn from the population is uniform (the same)

- A failure of each unit to have a uniform probability of being drawn from the population is known as **selection bias**
- Units are "selecting" into our data because they are more observable than other units

Random Sample

Definition

The probability of any given unit being drawn from the population is uniform (the same)

- A failure of each unit to have a uniform probability of being drawn from the population is known as **selection bias**
- Units are "selecting" into our data because they are more observable than other units
- Selection bias reduces our **generalizability** to the population because the data is not representative of the population

Random Sample

Can I randomly sample 10 students from this class to generalize to:

- the population of GW students?

Random Sample

Can I randomly sample 10 students from this class to generalize to:

- the population of GW students?
- the population of SMPA students?

Random Sample

Can I randomly sample 10 students from this class to generalize to:

- the population of GW students?
- the population of SMPA students?
- the population of Data Analysis students?

Exercise: Selection Bias

Controls: