

# Session 6: Working with data

---

Stats 60/Psych 10  
Ismael Lemhadri  
Summer 2020

# This time

---

- Types of data
- Scales of measurement
- Constructs vs. measurements
- Validity
- Reliability
- The ethics of data

# What is “data literacy”?

---

Language literacy



Data literacy

- Knowing how to find and prepare data for analysis
- Knowing how to analyze data
- Knowing how to interpret results from analyses

# What are data?

---

- data are plural
- you say “dah-ta”, I say “day-ta”
- A data set is composed of *variables*
  - Each variable contains information about some specific thing

# Types of data

Qualitative data: not composed of numbers

What is your favorite food?

Blueberries

Chocolate

Tamales

Pasta

Pasta

pizza

Mango

If you have programming experience, which programming languages do you have experience with?

None

Python

R

Java;  
Stata

Why are you taking this course?

It fulfills a General Education Breadth Requirement (WAY-AQR/WAY-FR)

It fulfills a degree plan requirement

It is not required but I am interested in the topic.

# Types of data

Quantitative data: data composed of numbers

Integers: counting numbers

The thought of being enrolled in a statistics course makes me nervous.

4                      7                      3                      7                      3                      6

Real numbers: numbers with decimals

How tall are you in inches?

74.0                      64.0                      65.0                      64.0                      64.0                      72.5

Binary values: True/False, 1/0

Have you ever taken a statistics course before?

No   No                      No                      No                      Yes                      Yes                      Yes                      Yes

# Qualitative data coded using numbers

---

- Discrete/categorical variables are often coded using numbers
  - 1:Republican, 2:Democrat, 3:Green
- Why would you want to do that?

# Recoding data

---

- We can also recode different types of numeric data
  - Binarizing data
    - Are you greater than 6 feet?
  - Rounding data
    - How tall are you (to the nearest inch)
  - Reversing the scale (often used for questionnaires)

Unfriendliness



Friendliness





# Scales of measurement

---

- The different values of a variable can differ by:
  - **Identity**
    - Each value of the variable has a unique meaning.
  - **Magnitude**
    - Values of the variable have an ordered relationship to one another. That is, some values are larger and some are smaller.
  - **Equal intervals**
    - Units along the scale of measurement are equal to one another. This means, for example, that the difference between 1 and 2 would be equal to the difference between 19 and 20.
  - **Absolute zero**
    - The scale has a meaningful zero point

# Nominal scale variables

---

- Each number represents a different thing
- Numbers do not actually represent quantities or ordering
  - 1=Republican, 2=Democrat, 3=Green, 4=Other
- What can we do with nominal variables?

Identity

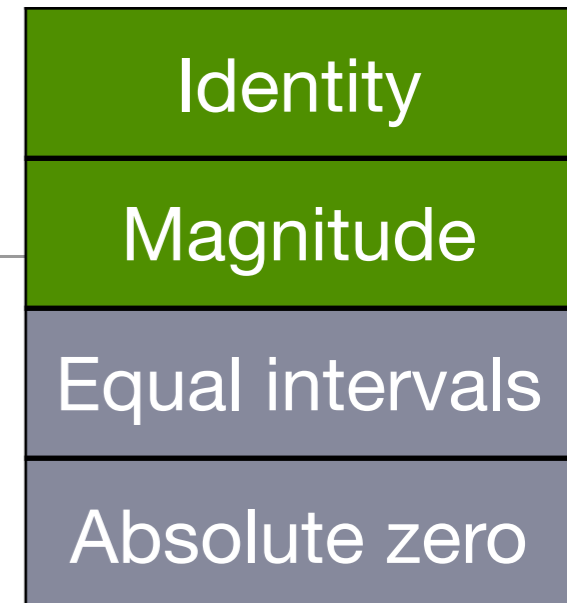
Magnitude

Equal intervals

Absolute zero

# Ordinal scale variables

- Values have an ordered relationship to one another
  - some values are larger and some are smaller
- Personality test items



	① Rarely/Never	② Occasionally	③ Often	④ Almost Always/Always
1 I plan tasks carefully.	①	②	③	④

# Interval scale variables

---

- The difference between values has a consistent meaning
- Temperature
  - Physical difference between 10 and 20 degrees is same as difference between 110 and 120 degrees

Identity

Magnitude

Equal intervals

Absolute zero

# Ratio scale variables

---

- Scale has a true zero
- Counts
  - Number of siblings
  - Balance in your checking account
- “Extensive” measurements
  - length or weight

Identity

Magnitude

Equal intervals

Absolute zero

# Some examples

---

- What type of scale does each of the following variables reflect?

# Measurement scale: Time to run a mile

Nominal

Ordinal

Interval

Ratio

# Measurement scale: Letter grades on a test (recoded as numbers)

Nominal

Ordinal

Interval

Ratio



# Measurement scale: License plate number

Nominal

Ordinal

Interval

Ratio

# Why does it matter whether a variable is measured on an ordinal, interval, or ratio scale?

We can only compute statistics on data that use a ratio scale.

**A**

We need to know the order of the values, which is only known for an ordinal scale

**B**

We can only plot the data if the variable is on a ratio scale

**C**

We can only apply certain arithmetic operations if the data are on an interval or ratio scale

**D**

# Why scales of measurement matter

---

- What kinds of operations can we perform?

	equal/not equal	Greater than/less than	Add/Subtract	Multiply/divide
Nominal				
Ordinal				
Interval				
Ratio				

# Why scales of measurement matter

---

- How can we summarize a variable?

	Mode	Median	Mean
Nominal			
Ordinal			
Interval			
Ratio			

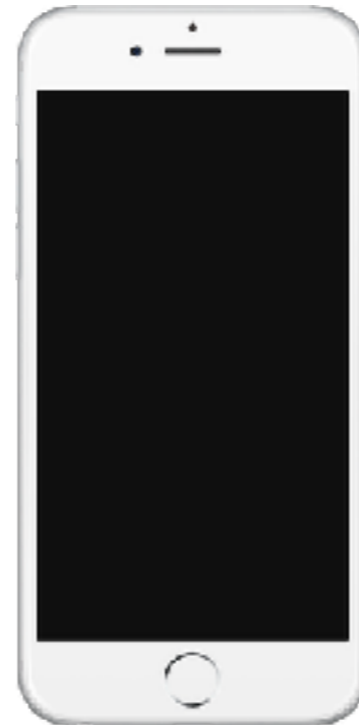
# Nonsensical results from misapplied statistics

---

iPhone 5



iPhone 6



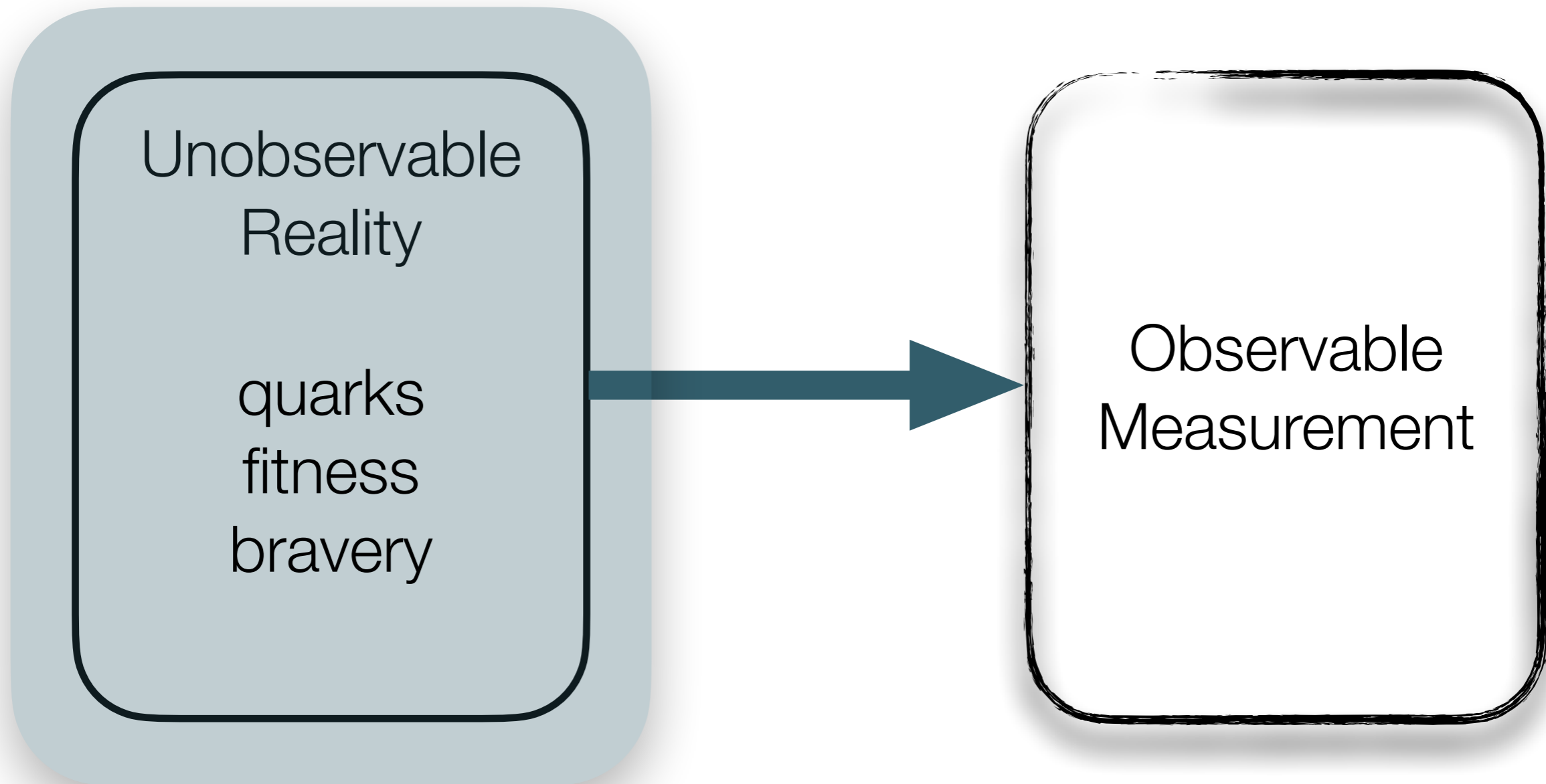
**iPhone 8**



My average iPhone model is 6.33

# Theoretical constructs vs measurements

---



# Some constructs and measurements

---

<b>Construct</b>	<b>Measurement</b>
Economic output	Gross domestic product
Speed of mental processing	Reaction time on a psychological test
Changes in Earth's climate over time	Temperature measurements from ocean buoys over time
Popularity of a person	Number of friends on Facebook

# Measurement error

---

- Rarely can we measure a construct without some error
- TPS:
  - Think of something that we regularly measure in the world. What are some of the possible causes of error in the measurement? If there are different ways to measure it, are those different ways subject to different causes of error?



# Measurements vary in their quality

---

- “Gold standard”
  - The most accurate way to measure something
  - Other methods are compared to this
- Sometimes we don't have a gold standard



A copy of the "provisional" metre installed 1796-1797, located in the wall of a building, 36 rue de Vaugirard, Paris.

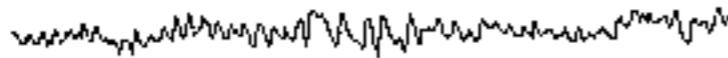
[https://en.wikipedia.org/wiki/History\\_of\\_the\\_metre](https://en.wikipedia.org/wiki/History_of_the_metre)

# Gold standard: Polysomnography

## Measuring sleep



Stage 1 -- Transition into sleep



Stage 2 -- Beginning of slow-wave sleep, marked by bursts of activity called sleep spindles



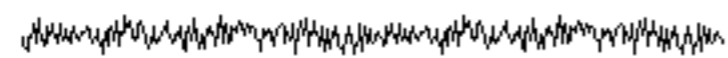
Stage 3-- Marked by appearance of very large, very slow waves called delta waves



Stage 4 -- Mostly delta waves



REM Sleep -- Fast, small waves indicating intense brain activity



## Alternative: Actigraphy



# The psychological construct of “impulsiveness”

---

- behavior without adequate forethought
- the tendency to act with less forethought than do most individuals of equal ability and knowledge
- a predisposition toward rapid, unplanned reactions to internal or external stimuli without regard to the negative consequences of these reactions
- Does this sound like a reasonable construct to measure?

# Barratt Impulsiveness Scale

---

	① Rarely/Never	② Occasionally	③ Often	④ Almost Always/Always
1 I plan tasks carefully.	①	②	③	④
2 I do things without thinking.	①	②	③	④
3 I make-up my mind quickly.	①	②	③	④
4 I am happy-go-lucky.	①	②	③	④
5 I don't "pay attention."	①	②	③	④
6 I have "racing" thoughts.	①	②	③	④

# Two aspects of measurement quality

---

- Validity
  - Does the test actually measure the construct that it claims to measure?
- Reliability
  - How well does the test measure the thing that it measures?



# Face validity

- Does the measurement seem on its face to answer the proper question?
  - Does the Barratt Impulsiveness Scale seem like it is a valid measure of the construct of “impulsiveness”?

	① Rarely/Never	② Occasionally	③ Often	④ Almost Always/Always
1 I plan tasks carefully.	①	②	③	④
2 I do things without thinking.	①	②	③	④
3 I make-up my mind quickly.	①	②	③	④
4 I am happy-go-lucky.	①	②	③	④
5 I don't “pay attention.”	①	②	③	④
6 I have “racing” thoughts.	①	②	③	④

# Construct validity

---

- Is the measurement related to other measurements in an appropriate way?

## Divergent validity

Impulsivness

I say things without thinking

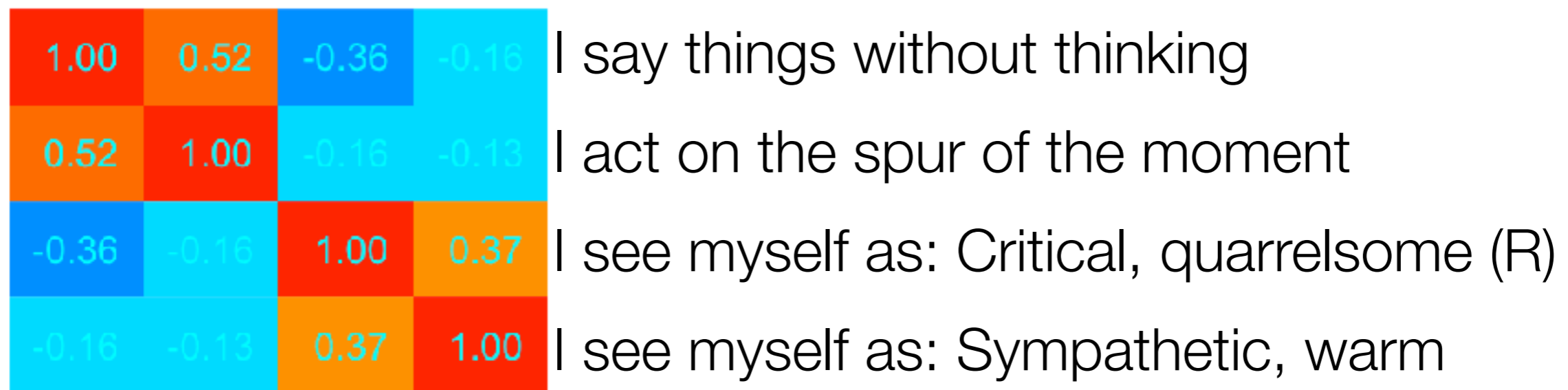
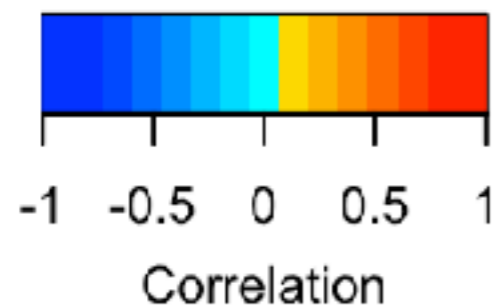
I act on the spur of the moment

Agreeableness

I see myself as: Critical, quarrelsome (Reversed)

I see myself as: Sympathetic, warm

## Convergent validity



Data from Eisenberg et al.

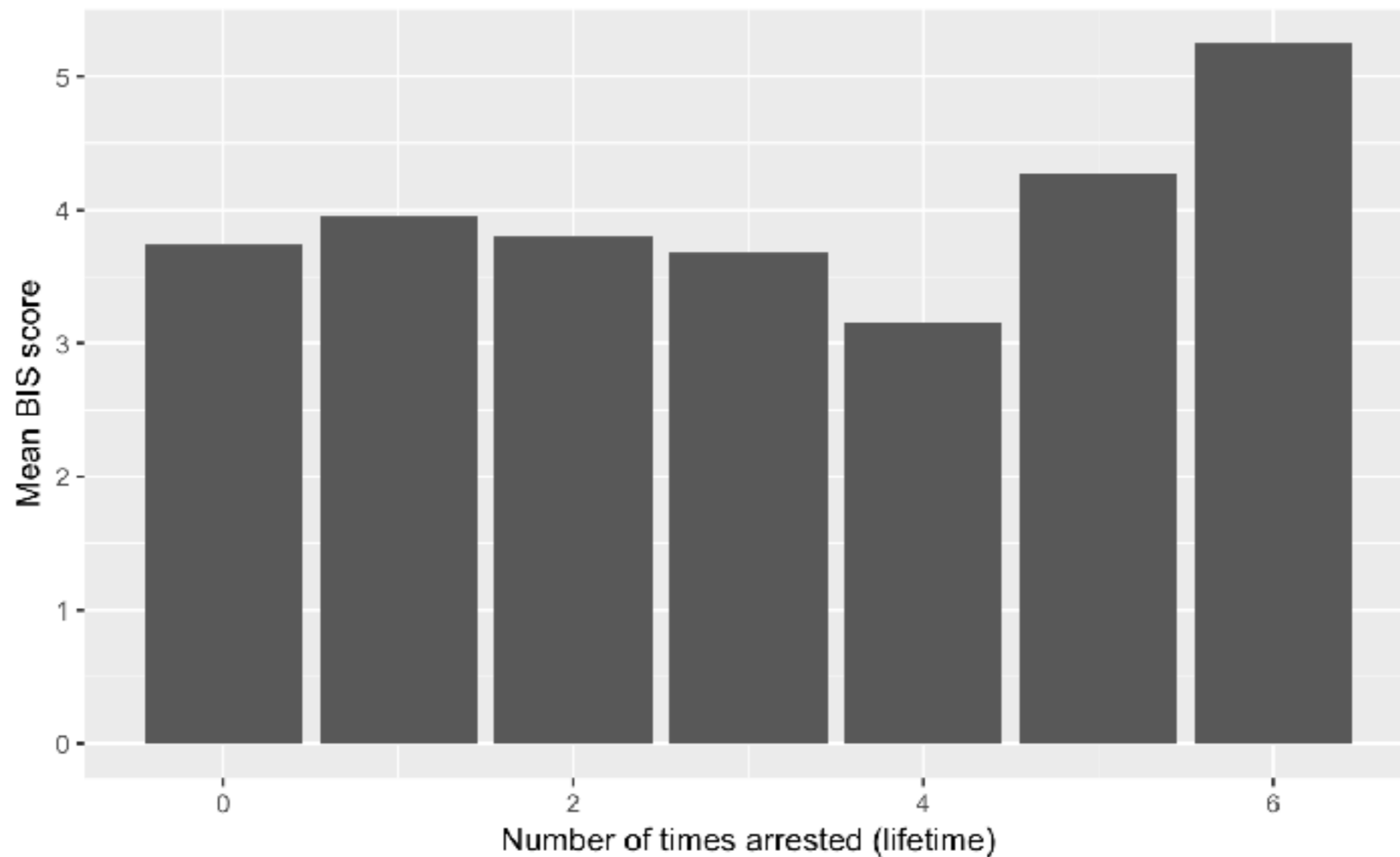
I say things without thinking  
 I act on the spur of the moment  
 I see myself as: Critical, quarrelsome (R)  
 I see myself as: Sympathetic, warm



# Predictive validity

---

- Is the measurement predictive of other outcomes?
  - What would you expect impulsiveness to predict?



# Group exercise

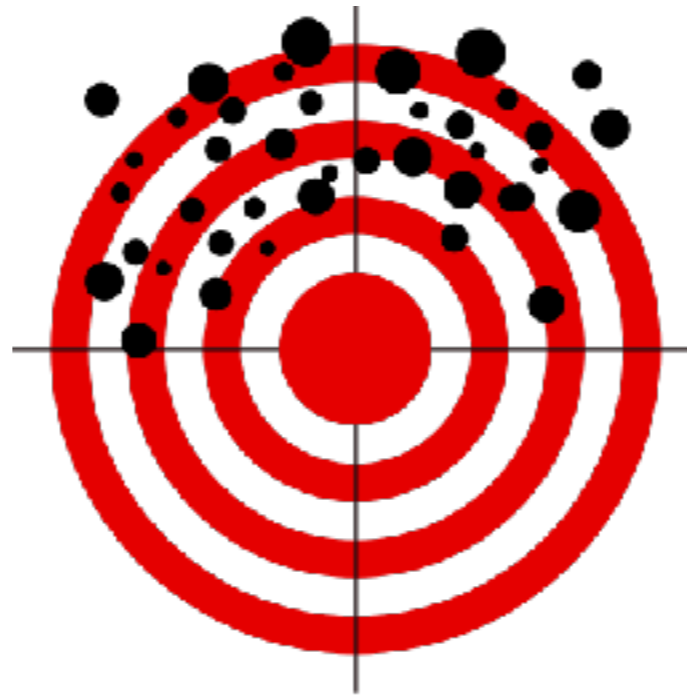
---

- My laboratory has developed a new 4-item survey to measure the construct of “fearlessness”. Here are the items (which the subjects rate on a 1-7 scale):
  - 1. I rarely avoid doing things I want to do because I am afraid of what might happen.
  - 2. I am often nervous in unfamiliar situations.
  - 3. I am often willing to do things that other people are afraid to do.
  - 4. When I am in an uncertain situation, I usually “go for it”
- What would you do to establish the validity of this survey? Address each of the types of validity.
- Would you need to recode any of the items?

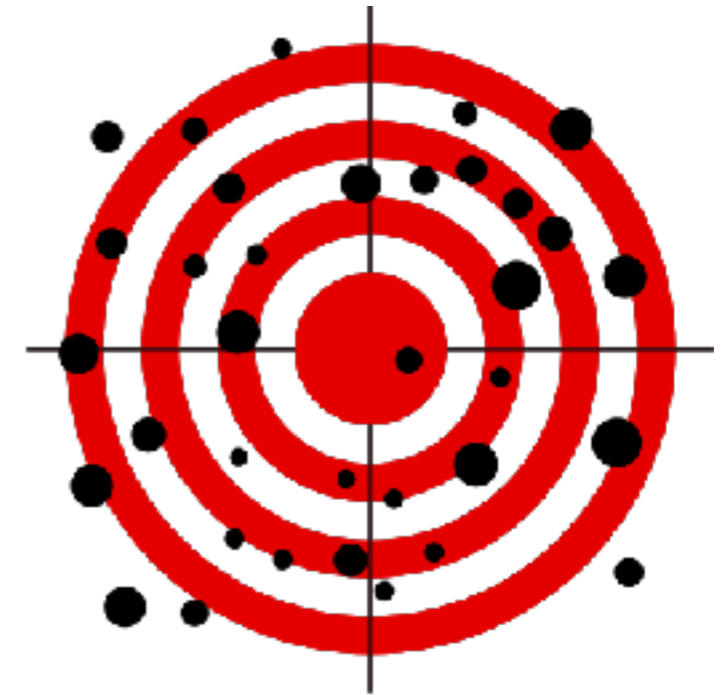
# Reliability

---

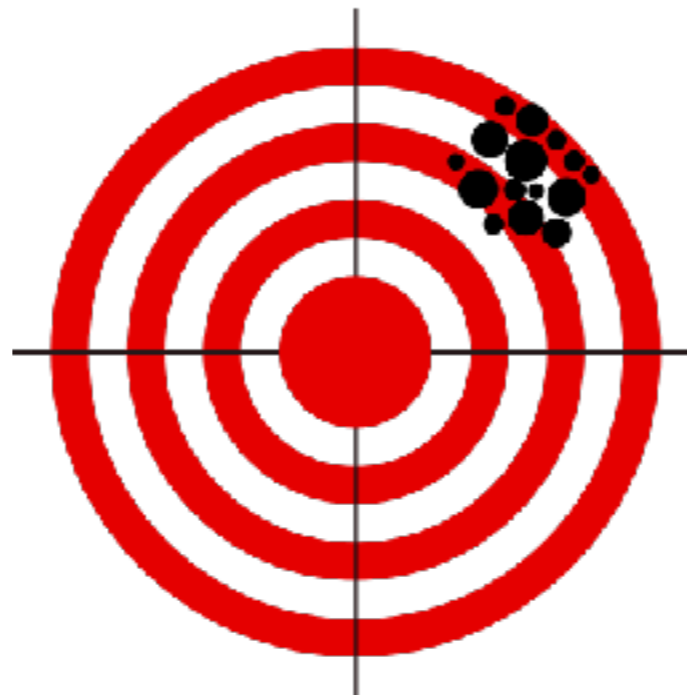
- Does the measure answer the question properly?



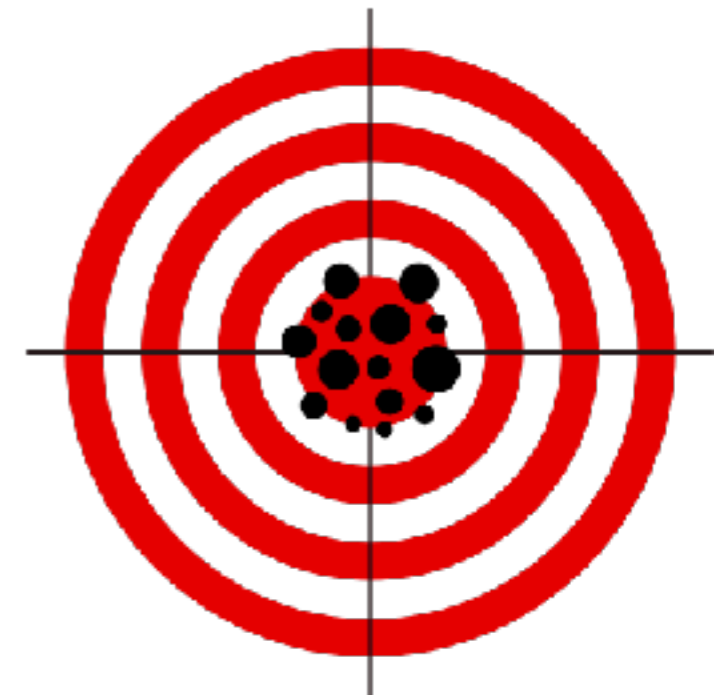
Unreliable & Invalid



Unreliable, But Valid



Reliable, Not Valid



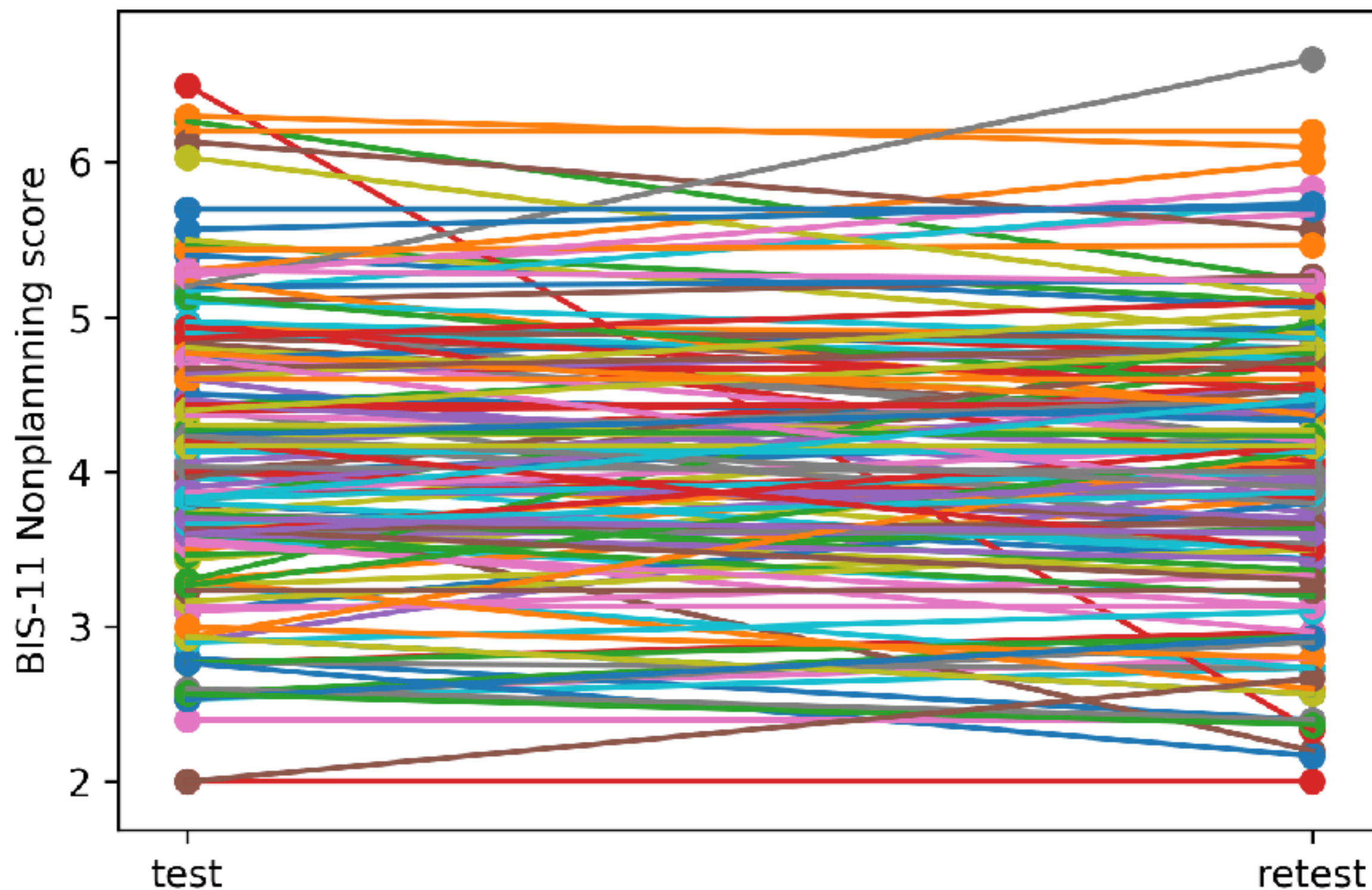
Both Reliable & Valid

# Test-retest reliability

---

- If we give the test multiple times, does it give the same answer?

# Test-retest reliability of BIS-11



# Inter-rater reliability

---

- If multiple people rate the same thing, do we get the same results?
- What factors do you think would influence inter-rater reliability?

# The ethics of data

---

## **The Stanford Daily**

### **Privacy breaches in University file system affect 200 people**

November 17, 2017

Stanford is in the process of notifying some 200 people — a mix of employees and former students — that their privacy may have been breached due to incorrect settings in one of the University's file-sharing systems.

Until this week, files including sexual violence records based on counseling sessions, confidential University statistics and emails to the Office of Judicial Affairs — some with names and email addresses attached — were left broadly available on an internet server that students, faculty and staff from over 50 institutions regularly use. Any Stanford faculty, student or staff member with a SUNet ID was able to access the sensitive files; The Daily also found that an MIT student username and password were able to grant access.

# Identifiability

---

- The Health Insurance Portability and Accountability Act of 1996 (HIPAA) defines 18 unique identifiers for human data
  - Name
  - Address (all geographic subdivisions smaller than state, including street address, city county, and zip code)
  - All elements (except years) of dates related to an individual (including birthdate, admission date, discharge date, date of death, and exact age if over 89)
  - Telephone numbers
  - Fax number
  - Email address
  - Social Security Number
  - Medical record number
  - Health plan beneficiary number
  - Account number
  - Certificate or licence number
  - Any vehicle or other device serial number
  - Web URL
  - Internet Protocol (IP) Address
  - Finger or voice print
  - Photographic image -  
Photographic images are not limited to images of the face.
  - Any other characteristic that could uniquely identify the individual

Do you think those are sufficient?



# Data privacy

---

- how much do you worry about your privacy?
- Let's say that I presented data from a hypothetical student:
  - Major: IR
  - Year 4
  - Height: 74 inches
  - Gender: Female
  - Favorite food: Sushi
  - House number: 2514
  - Anxious, easily upset: 7
  - The thought of being in a stats class makes me nervous: 7
  - Dealing with numbers makes me uneasy: 7
- How hard would it be to identify this person?
- If it was you, would you be upset if your data were released?

# Recap

---

- Variables have different scales of measurement
- The scale of measurement determines what one can do with the data
- Some variables are meant to measure unobservable constructs
- Validity: Is it measuring the right thing?
- Reliability: Is it measuring the thing right?
- With data comes great power, and thus data thus require great care