Research Methods in Cognitive Science

# Week 5: Measurement

Jason Geller, Ph.D

2021-09-28

# Housekeeping

- Next Tuesday: A120 (RuCCS Eye-tracking lab)
    - Team 1 - 3:00-3:30
    - Team 2 - 3:30: 4:00
- Next Thursday
    - Sarah Colby (University of Iowa)
- Tuesday (October 12th)
    - Team 3 - 3:00-3:30
    - Team 4 - 3:30-4:00
    - Team 5 - 4:00-4:30

## Last Class

- 680 422

05:00

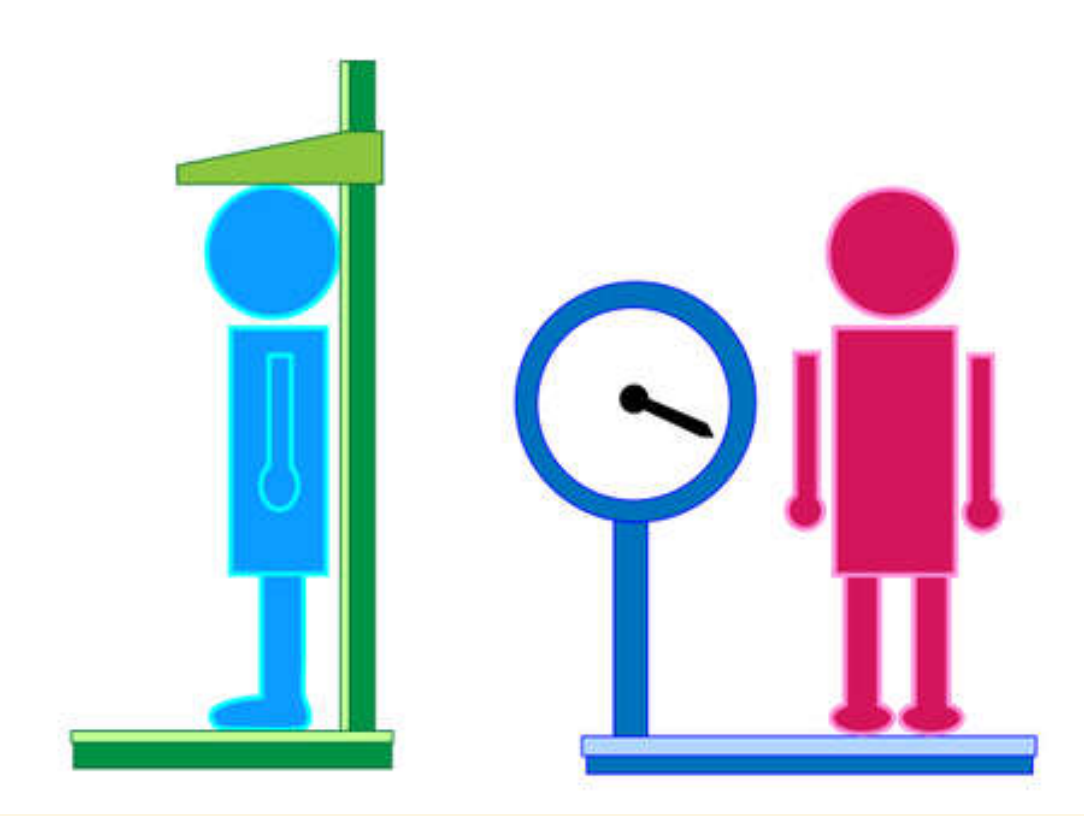
# Today

- What is measurement?
- Scales of measurement
- Reliability and validity
- Measurement in practice: listening effort

> "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality." -Edward L. Thorndike

# Measurement

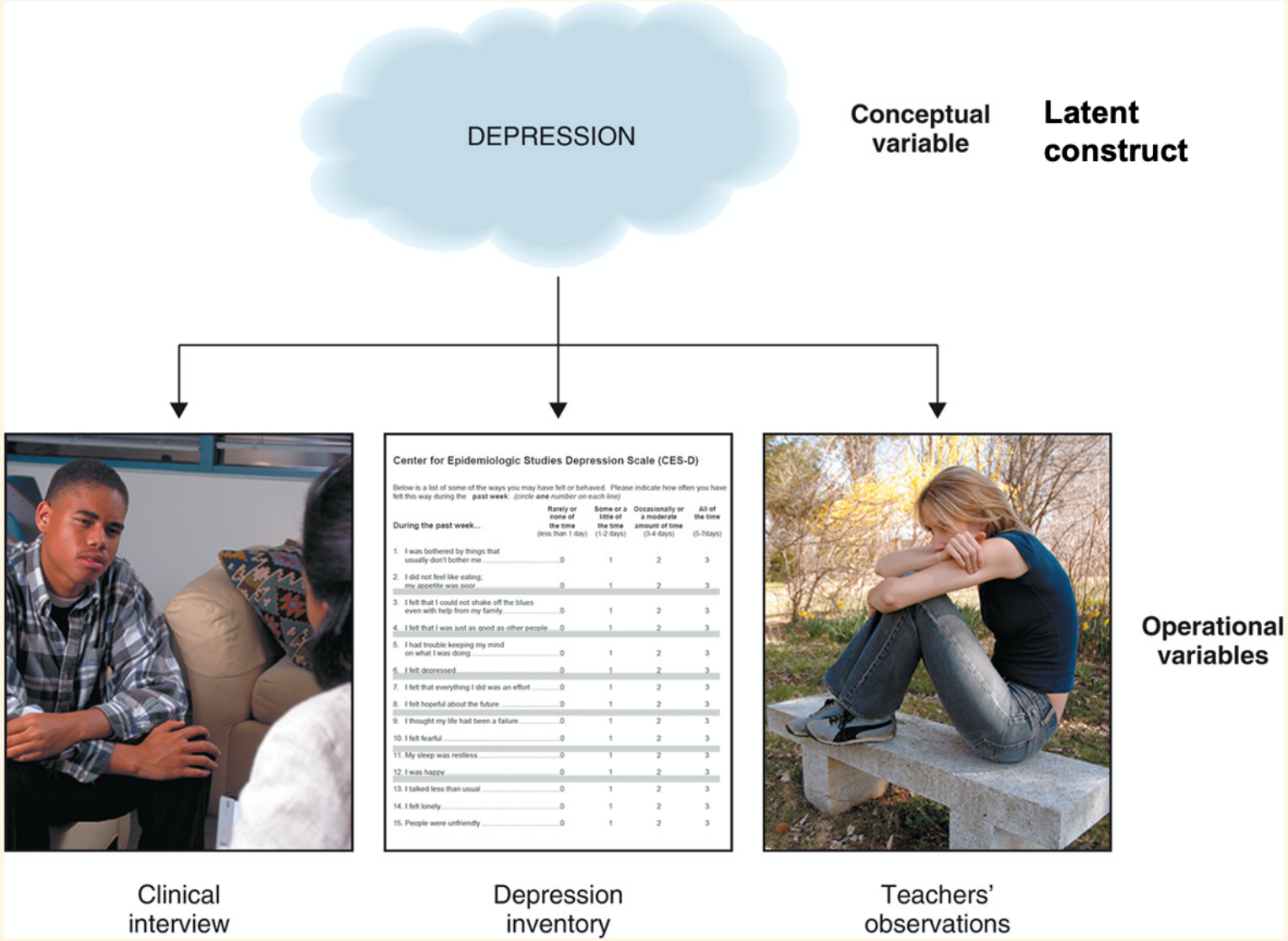
> The assignment of scores so that the scores represent some characteristic of the individuals



# What things do we want to measure?

- Depression
- Effort
- Intelligence
- Memory
- Social support
- Extroversion
- Eating behavior
- Parent child relationships
- Attention
- Burn out
- Hopelessness

# Constructs ≠ Variables



# Scales of Measurement

- Variables are defined and categorized four ways:

1. Nominal -> categorical data
2. Ordinal
3. Interval
4. Ratio

**NOIR**

# Nominal

- Nominal ≈ name, so numbers on a nominal scale just name or stand for a category or individual

- Numerals arbitrarily assigned to name events/objects

– Given 2 nominal measurements:

- Can determine whether the same or not
- Not able to tell if one has more or less of measured attribute

- Examples: gender, experimental vs control group

# Ordinal

– Data have characteristics of nominal scale + more

– Numbers indicate the ordering of individuals on some dimension

- Given 2 ordinal measurements:
    - Can state whether they have equal amounts of the attribute or not
    - May not be able to make statements about the difference between the pair of scores
- Examples: Ranking on a task, rating your preferences

# An Olympic Example

| Rank by Gold | Country | # of Athletes | # of Gold Medals |
|---|---|---|---|
| 1 | USA | 1,200 | 46 |
| 2 | People's Republic of China | 600 | 38 |
| 3 | Great Britain | 1,500 | 29 |

# Interval

- Data have characteristics of ordinal scale + more
- Intervals have consistent meaning – equal units
- No true zero
- Example: temperature in °F

# Ratio

- Intervals have consistent meaning – equal units

– True zero

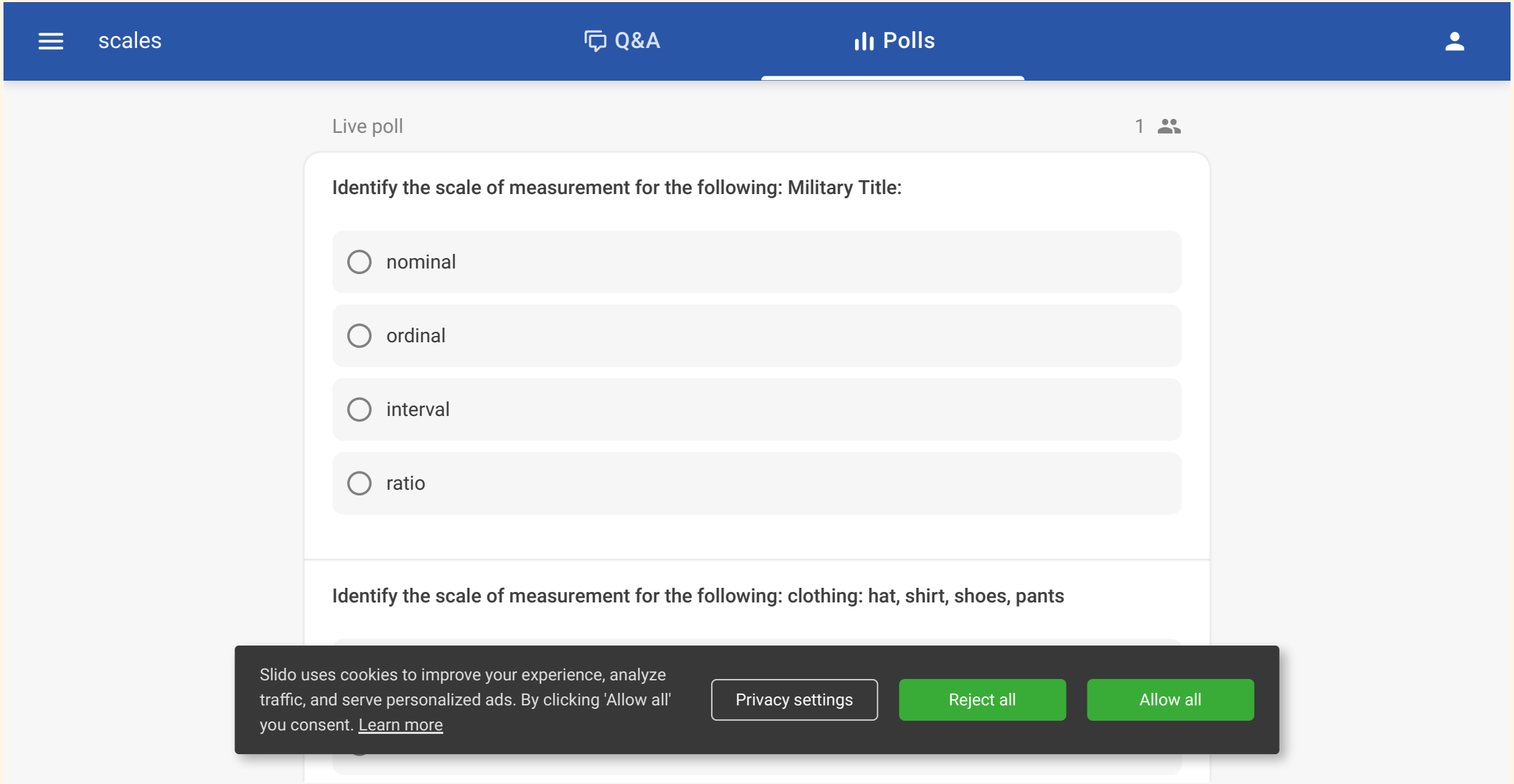
– Ratios are meaningful

– Examples: frequencies, times, rates

# An Olympic Example

| Rank by Gold | Country | # of Athletes | # of Gold Medals |
|---|---|---|---|
| 1 | USA | 1,200 | 46 |
| 2 | People's Republic of China | 600 | 38 |
| 3 | Great Britain | 1,500 | 29 |

# Knowledge Check



# Reliability

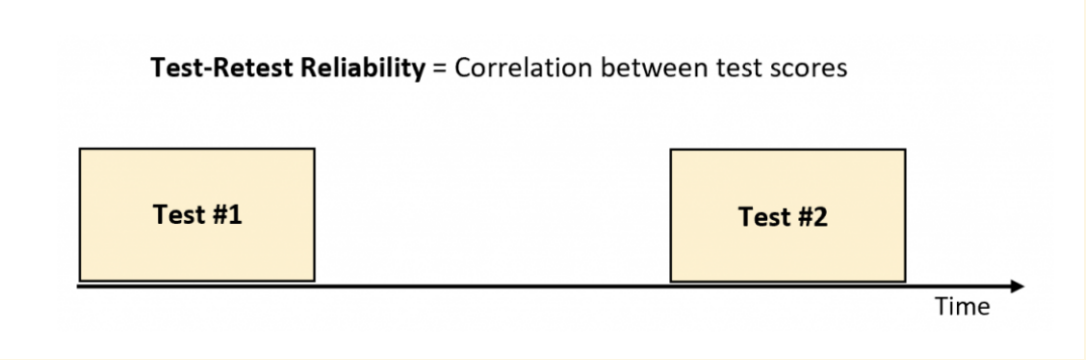
> How consistent or how precise a measure/method is

- Test-Retest (over time)
- Internal (across time)
- Interrater (between different researchers)

# Reliability

## Consistency of a measure:

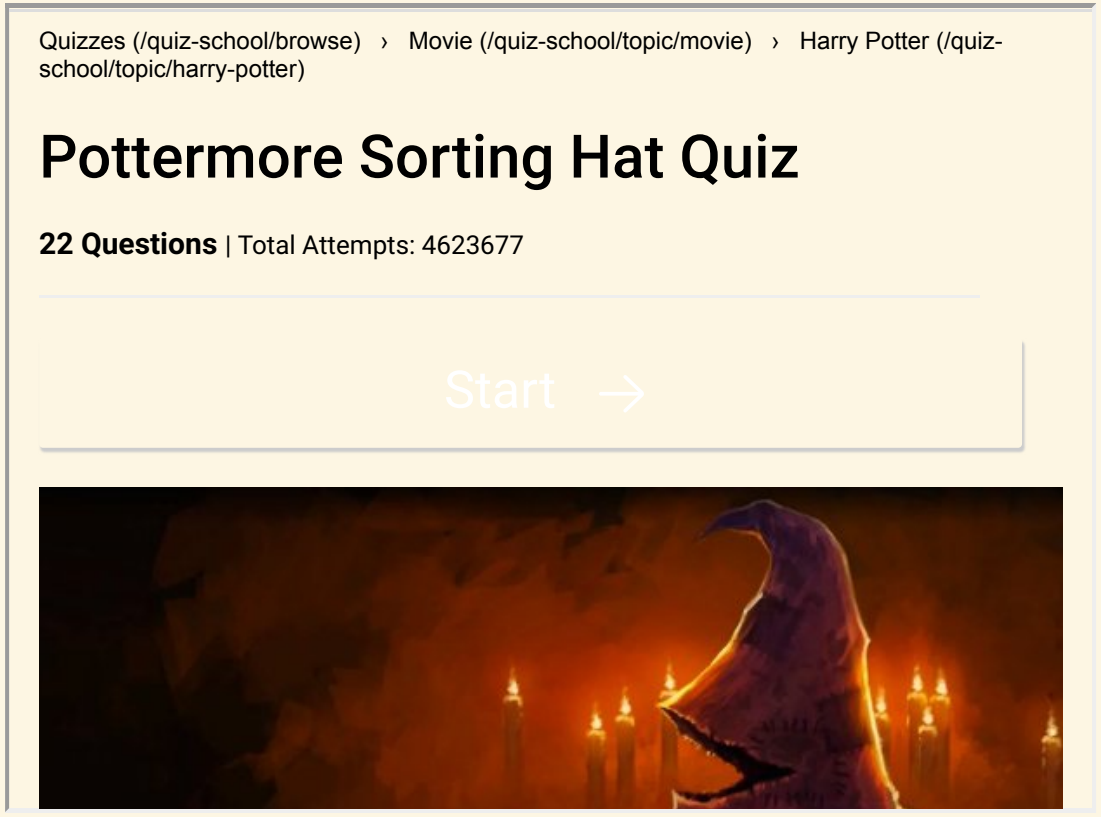
- Test-Retest (across time)



# Reliability

## Consistency of a measure:

- Test-retest (across time)



# Reliability

> Consistency of a measure:

- Internal (across items)

1. I love Halloween. **Agree**
2. I feel happy when I decorate my house for Halloween. **Agree**
3. I feel angry when the Halloween season is approaching. **Disagree**

    - Cronbach's $\alpha$
        - .8

# Reliability

Consistency of a measure:

- Interrater (across different researchers)

# Validty

> Accuracy [e.g., are we really measuring what we think we are measuring?]

- Face validity

  > The extent to which a measurement method appears “on its face” to measure the construct of interest

- Criterion or convergent validity

  > The extent to which people’s scores on a measure are correlated with other variables (known as criteria) that one would expect them to be correlated with.

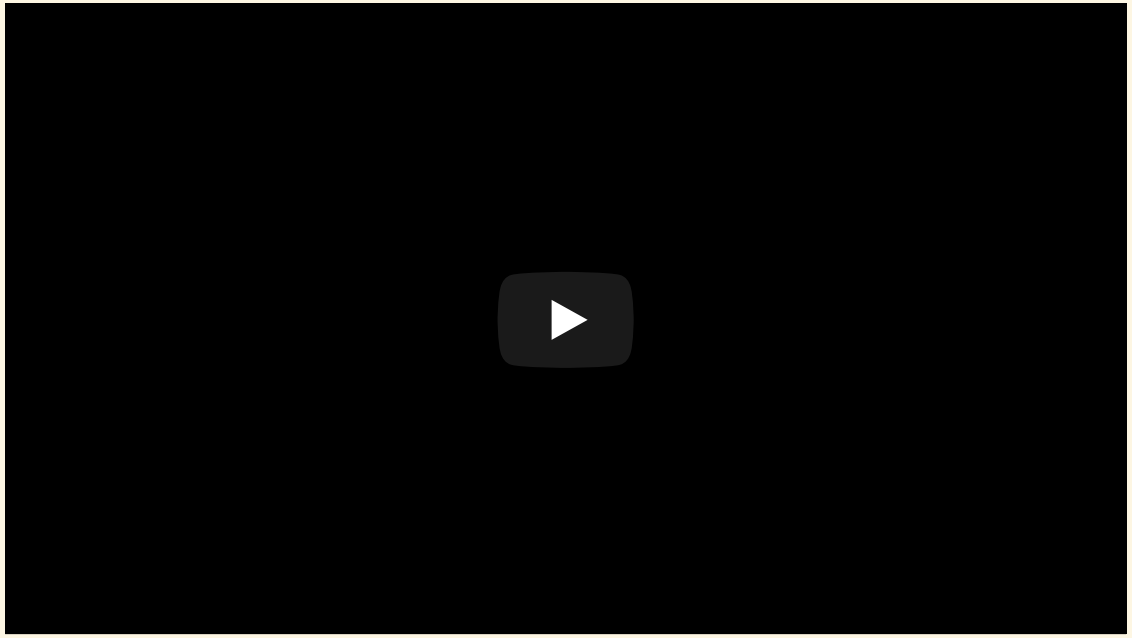
- Discriminate or divergent validity

  > The extent to which scores on a measure of a construct are not correlated with measures of other, conceptually distinct, constructs and thus discriminate between them.

# Bank Robbery

- Eyewitness memory plays an important role in helping police solve crimes. However, people's abilities to accurately recall what they saw can substantially impact whether a criminal is convicted—and equally, if an innocent person is wrongfully convicted. So, it's important to get it right.

- First, let's find out how well you can remember what happens during a bank robbery



# Viewing #1

- You have now viewed a clip of a simulated bank robbery.

Take a few minutes to write down your description of the main offender

# Work with partner

Take a few minutes to write down your description of the main offender

- Inter-observer agreement = `# agreements X 100 / # agreements + # disagreements` For example: if there were 5 agreements and 3 disagreements... [a] 5 x 100 = 500 [b] 5 + 3 = 8 [c] 500/8 = 62.5% inter-observer agreement

- What percentage agreement did you end up with? What do you think it says about the reliability of the instructions you were given?

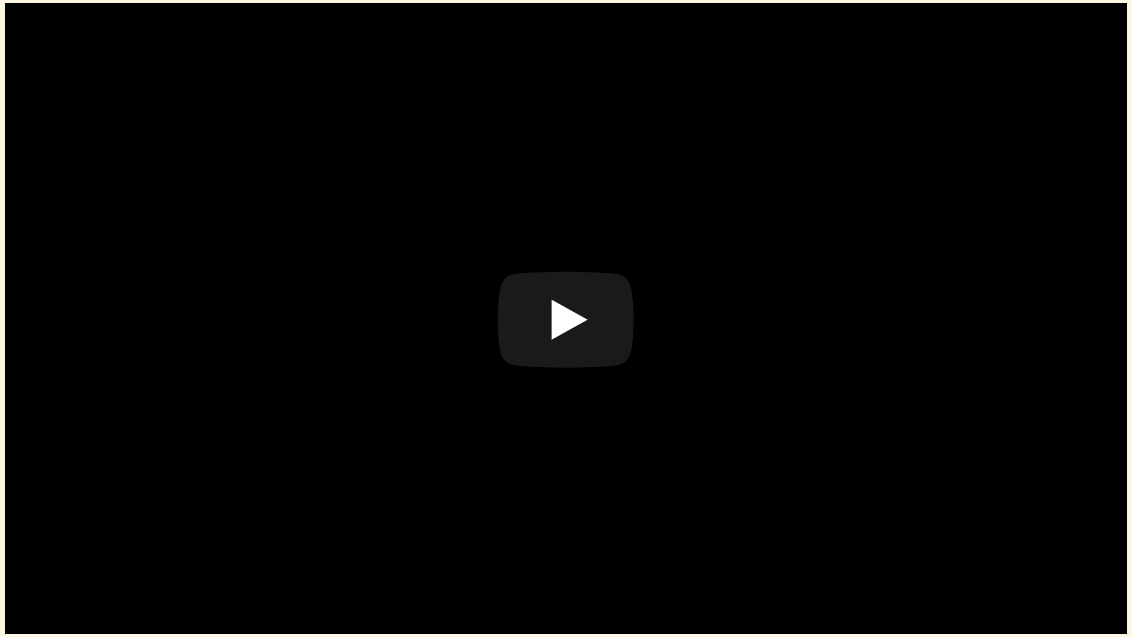
05:00

# Group Discussion

- What sort of percentage agreements did we get?
- What do you think these percentages tell us about the reliability of the instructions you were given?

# Viewing #2

- We will watch the video again.
- Then you will get new instructions for describing the offender.
- Then you will work with your partner to calculate inter-rater reliability again



# Viewing #2

Using the checklist below, describe the main offender:

- Were they male or female?
- What was their hair colour?
- What was their skin colour?
- What colour were their eyes?
- What was the colour of their shirt?
- Were they wearing a jacket? If so, what colour was it?
- Were they wearing long pants, jeans, or shorts?
- What colour was their pants/jeans/shorts?
- Were they wearing glasses?
- Were they wearing a hat?
- Were they wearing a balaclava?
- Did he/she have a gun?
- Did he/she have a knife?
- Were they carrying anything? If yes, what was it?

# Work with partner

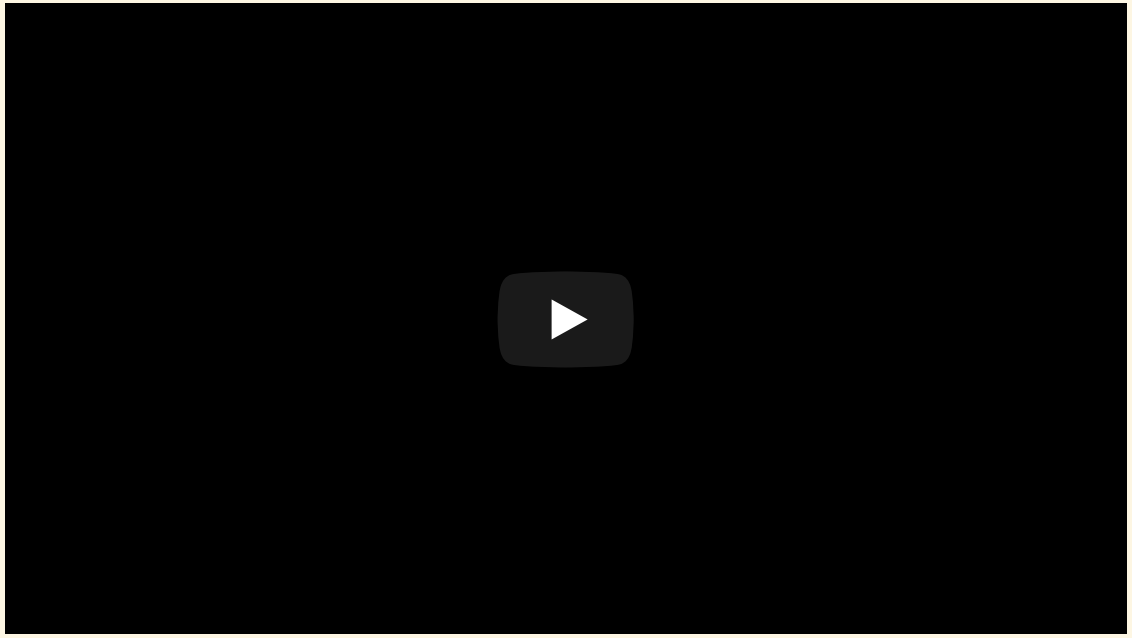
Take a few minutes to write down your description of the main offender

- Inter-observer agreement = `# agreements X 100 / # agreements + # disagreements`
- What percentage agreement did you end up with? What do you think it says about the reliability of the instructions you were given?

> For example: if there were 5 agreements and 3 disagreements [a] 5 x 100 = 500 [b] 5 + 3 = 8 [c] 500/8 = 62.5% inter-observer agreement

05:00

# Viewing #3



Using the checklist below, describe the main offender

- Were they male or female?
- What was their hair colour?
- What was their skin colour?
- What colour were their eyes?
- What was the colour of their shirt?
- Were they wearing a jacket? If so, what colour was it?
- Were they wearing long pants, jeans, or shorts?
- What colour was their pants/jeans/shorts?
- Were they wearing glasses?
- Were they wearing a hat?
- Were they wearing a balaclava?
- Did he/she have a gun?
- Did he/she have a knife?
- Were they carrying anything? If yes, what was it?

# On your own this time

Now, using the same formula, calculate how well you agree within yourself (test-retest reliability). That is, what is the level of correspondence between your observations at Viewing #2 and your observations at Viewing #3?

- Inter-observer agreement = `# agreements X 100 / # agreements + # disagreements`

05:00

# Group Discussion

How do these results compare with the results for the inter-rater reliability calculations, i.e., are they different? In what way? Any ideas why or why not?

# Listening Effort

- Conceptual definition

> The deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a listening task

| Accuracy | Effort |
|---|---|
| HIGH | LOW |

| Accuracy | Effort |
|---|---|
| HIGH | HIGH |

# Operationalization

- How do we measure "accuracy"?

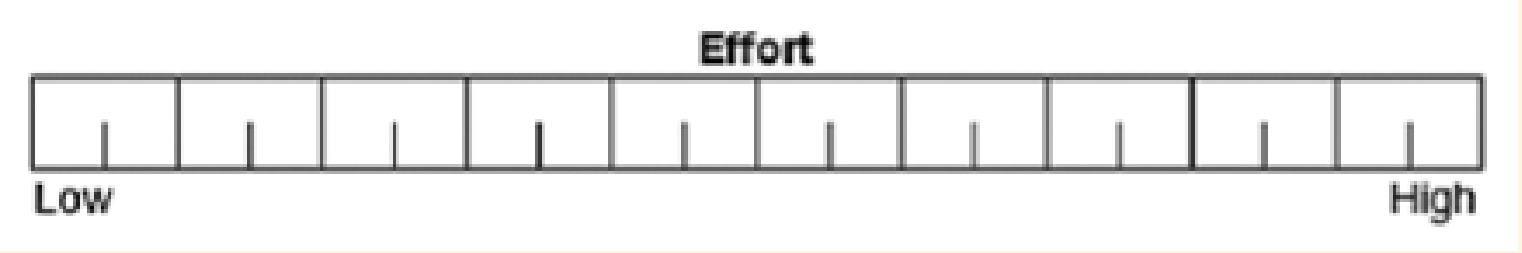
Hear:
"he killed the dragon with his sword"

6/7 words

# Operationalization

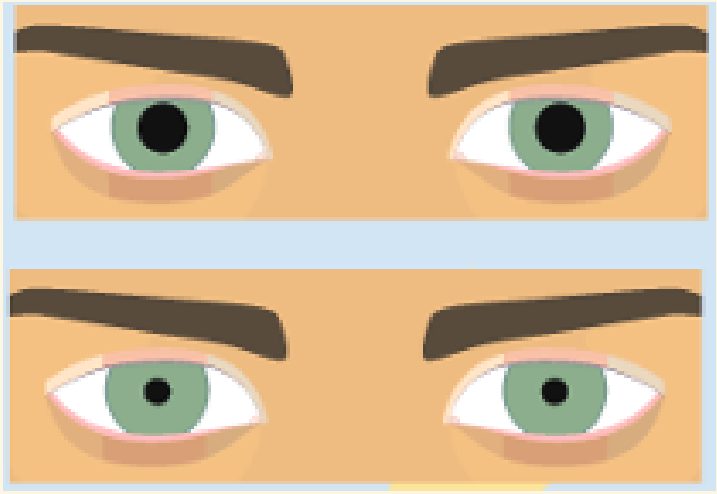
- How do we measure "effort"
    - Self-report:

“how hard did you have to work mentally to accomplish your level of performance?”



# How do we measure "effort"

- Physiological measures
  - Pupil size
  - Skin conductance (GSR)
  - Heart rate



# How do we measure "effort"

- Behavioral measures
  - Recall

Hear:
cat. . .table . . . fish . . .
memory . . .
danger . . . apple . . .
Nevada . . . jumping . . .
. telescope

DV: % words recalled. More effort ➔ fewer words recalled

# How do we measure "effort"?

- Behavioral measures
  - dual-task

Hear:
cat. . .table . . . fish . . .
memory . . .
danger . . . apple . . .
Nevada . . . jumping . . .
. telescope

Task: press a button as quickly as possible when you hear nouns

DV: time to noun judgment. More effort ➔ slower responses

# How many measures of listening effort in the literature?

- 24!

# Measuring Listening Effort (Strand et al., 2018)

# Jingle and Jangle Fallacy (Thorndike, 1904; Kelley, 1927; Flake & Fried, 2020)

- Jingle

> Falsely assuming that two tasks measure the same construct because they have the same name

# Measuring Listening Effort (Strand et al., 2018)

# Jingle and Jangle Fallacy (Thorndike, 1904; Kelley, 1927; Flake & Fried, 2020)

- Jingle

> Falsely assuming that two tasks measure the same construct because they have the same name

- Jangle

> Falsely assuming that two tasks measure different constructs because they have different names

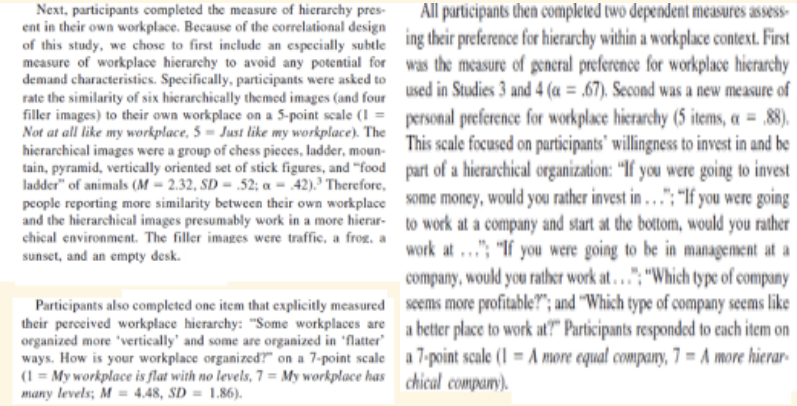
# Key Points

- Constructs ≠ variables
- Constructs are measured multiple ways, which may lead to different outcomes
- Thinking carefully about measurement is fundamental to understanding replication

- Replication Crisis [2011-2017]
- Theory Crisis [2017-2025]
- Falsification Crisis [2025-2030]
- Measurement Crisis [2030-2036]
- Collaboration Crisis [2036-2048]
- Golden Age of Psychology [2050-now]
- Discovery of Pre-Cognition

# Group Work

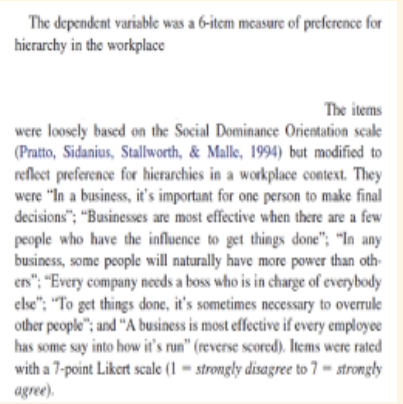
- What are the two key constructs of interest (across the three studies)?
- How were the constructs measured?
- How did the authors select the measures and where do the measures come from?
- How were the measurement decisions in these studies described and justified?



Next, participants completed the measure of hierarchy present in their own workplace. Because of the correlational design of this study, we chose to first include an especially subtle measure of workplace hierarchy to avoid any potential for demand characteristics. Specifically, participants were asked to rate the similarity of six hierarchically themed images (and four filler images) to their own workplace on a 5-point scale (1 = *Not at all like my workplace*, 5 = *Just like my workplace*). The hierarchical images were a group of chess pieces, ladder, mountain, pyramid, vertically oriented set of stick figures, and "food ladder" of animals ($M = 2.32$, $SD = .52$; $\alpha = .42$).[3] Therefore, people reporting more similarity between their own workplace and the hierarchical images presumably work in a more hierarchical environment. The filler images were traffic, a frog, a sunset, and an empty desk.

Participants also completed one item that explicitly measured their perceived workplace hierarchy: "Some workplaces are organized more 'vertically' and some are organized in 'flatter' ways. How is your workplace organized?" on a 7-point scale (1 = *My workplace is flat with no levels*, 7 = *My workplace has many levels*; $M = 4.48$, $SD = 1.86$).

All participants then completed two dependent measures assessing their preference for hierarchy within a workplace context. First was the measure of general preference for workplace hierarchy used in Studies 3 and 4 ($\alpha = .67$). Second was a new measure of personal preference for workplace hierarchy (5 items, $\alpha = .88$). This scale focused on participants' willingness to invest in and be part of a hierarchical organization: "If you were going to invest some money, would you rather invest in . . ."; "If you were going to work at a company and start at the bottom, would you rather work at . . ."; "If you were going to be in management at a company, would you rather work at . . ."; "Which type of company seems more profitable?"; and "Which type of company seems like a better place to work at?" Participants responded to each item on a 7-point scale (1 = *A more equal company*, 7 = *A more hierarchical company*).



The dependent variable was a 6-item measure of preference for hierarchy in the workplace

The items were loosely based on the Social Dominance Orientation scale (Pratto, Sidanius, Stallworth, & Malle, 1994) but modified to reflect preference for hierarchies in a workplace context. They were "In a business, it's important for one person to make final decisions"; "Businesses are most effective when there are a few people who have the influence to get things done"; "In any business, some people will naturally have more power than others"; "Every company needs a boss who is in charge of everybody else"; "To get things done, it's sometimes necessary to overrule other people"; and "A business is most effective if every employee has some say into how it's run" (reverse scored). Items were rated with a 7-point Likert scale (1 = *strongly disagree* to 7 = *strongly agree*).