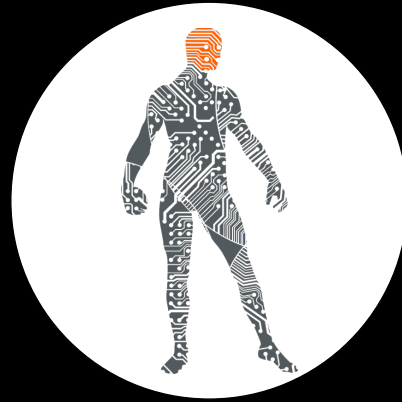


RECAP



Aalto-yliopisto
Aalto-universitetet
Aalto University



Research Methods in Engineering Psychology – Lecture 3

B.Sc. Engineering Psychology
Prof. Dr. Robin Welsch

Modules today



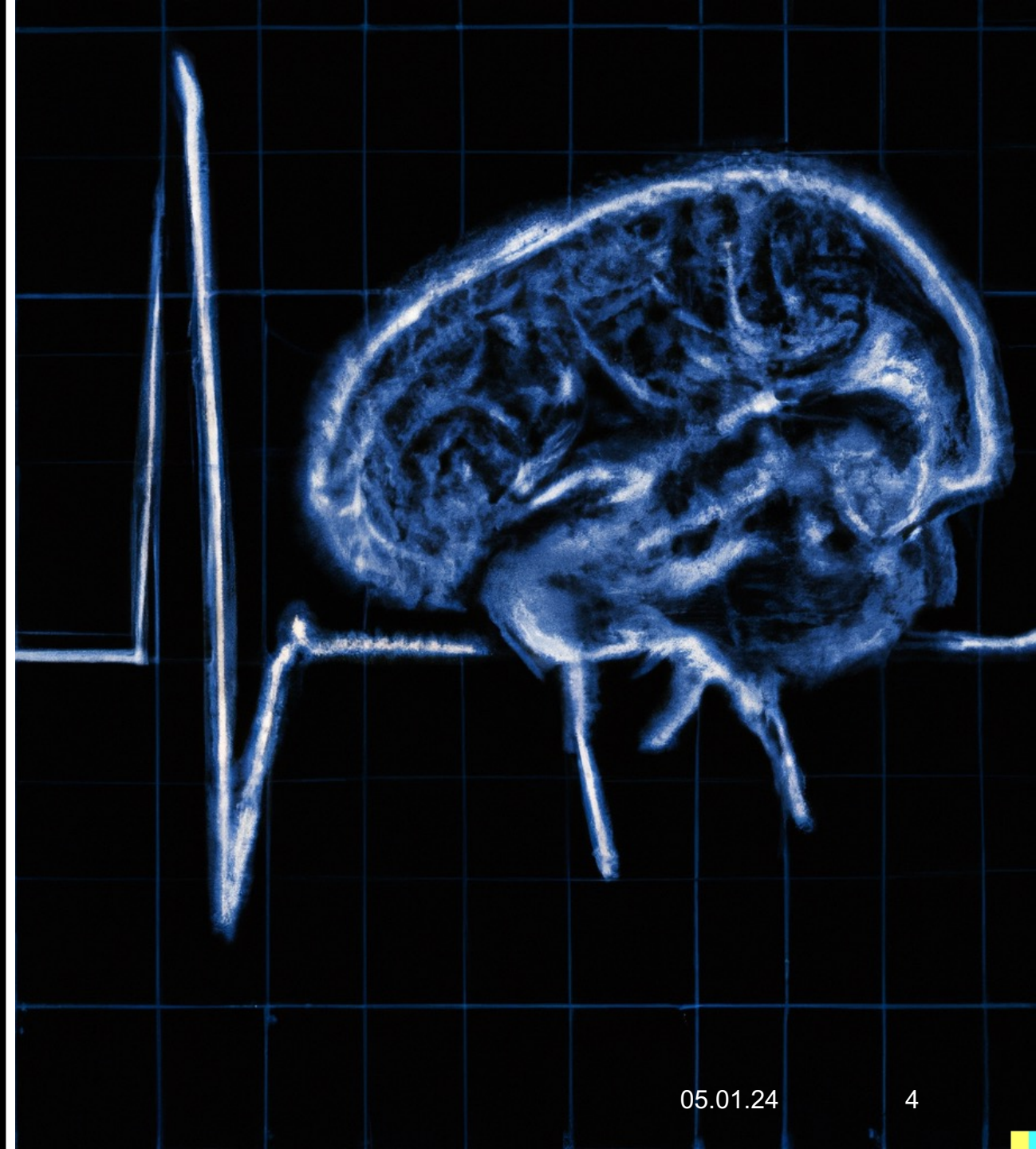
**Measurement in
Psychology**



**Evaluating
empirical research**

Measurements in Psychology

- Types of measurements
- Steeven's theory of measurement
- Operationalization
- Psychological scales
- Common research measures



How can we quantify this?



<https://www.youtube.com/shorts/Y74D3DKSFhw>
<https://www.youtube.com/watch?v=ZdOj7B0pcpE>
<https://www.youtube.com/watch?v=JHQ8UAjoVVc>

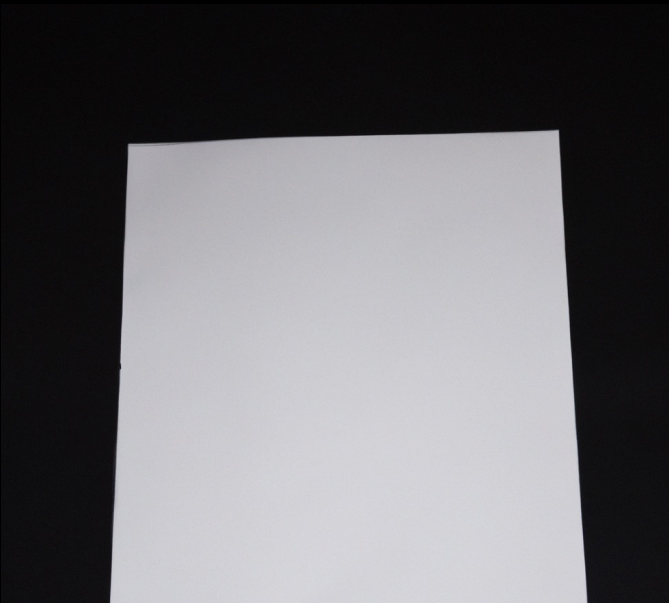
**How do measure what people
experience, feel or do?**

Making psychological constructs measurable

- Some things can be directly measured, e.g. body height
- Other psychological constructs e.g, love, are useful but cannot be directly observed.
- Operationalization: Process of making abstract concepts measurable, e.g. buying lots of flowers signals love → number of flowers
- Psychological measurement: A systematic procedure for assigning observations to categories or values of psychologically relevant variables, e.g. counting the number of flowers you bought

Types of quantitative psychological measurement

Subjective measures



Behavioral measures



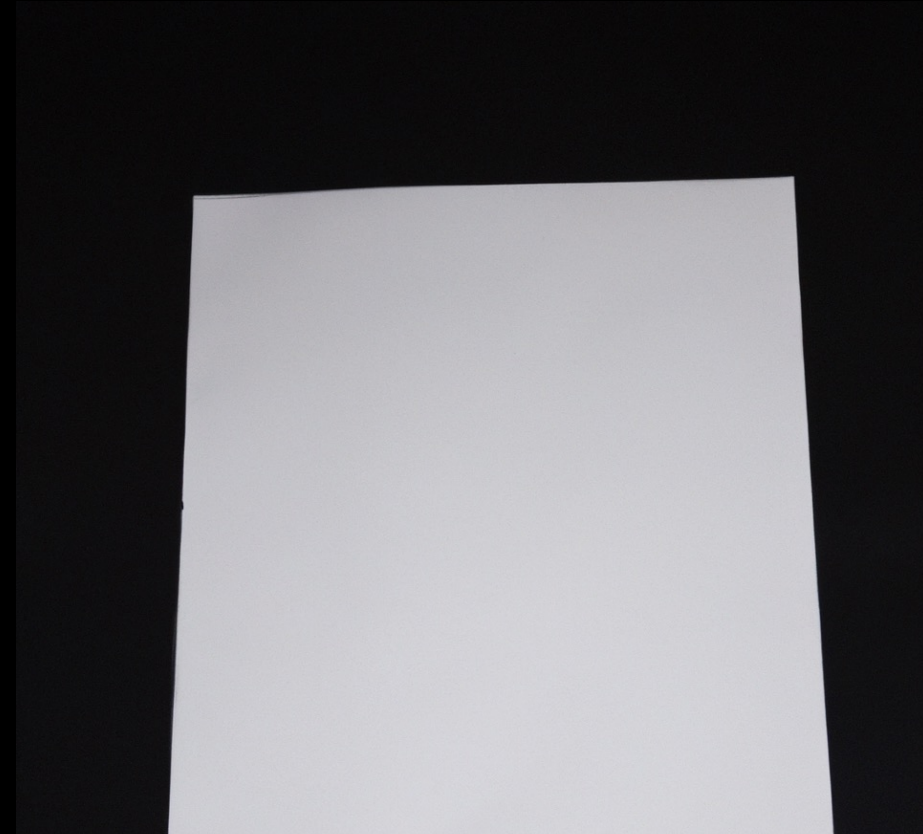
Psychophysiology



Subjective measures

Any measure that relies on self-report or subjective evaluation

- In engineering psychology, subjective measures are techniques used to assess people's perceptions, attitudes, and subjective experiences
- Typical measures are:
 - **Self-report measures: Asking people to report on their attitudes, preferences, or experiences with a product or task.**
 - Focus groups: Gathering a small group of people to discuss and provide feedback on a product or task.
Interviews: Asking people in-depth questions about their experiences or attitudes towards a product or task.
 - Critical incident technique: Asking people to describe specific instances or events related to their use of a product or task.
 - Think-aloud protocol: Asking people to verbalize their thoughts and processes as they perform a task.



Types of questions in psychological research

- **Open-ended questions:** do not have a fixed set of responses, and allow participants to provide their own answers in their own words.
- **Closed-ended questions:** have a fixed set of responses that participants must choose from.
- **Forced-choice questions:** closed-ended questions that require participants to choose between two or more options, even if none of the options accurately reflect their views or experiences.
- **Rating scale questions:** questions that ask participants to rate their responses on a scale, such as a Likert scale or a semantic differential scale.
- **Projective questions:** Questions that ask participants to respond to ambiguous stimuli, such as pictures or stories, in order to reveal thoughts or feelings. → not used anymore due to problems of reliability and validity
- An item is a specific question or task that is included in a research study or survey.

Psychological scales

Packaging questions together

- Psychological scales are standardized tools used to measure psychological constructs such as personality, intelligence, or attitudes
- There are many different types of psychological scales, e.g.:
 - Self-report scales require the individual to answer questions or make statements about themselves (e.g. BFI: Big Five inventory)
 - Observer-report scales require someone else to rate the individual (attachment style in children)
 - Performance-based scales measure the individual's behavior or abilities in a specific task or situation (Wechsler Adult Intelligence Scale)
- It is important to consider the reliability and validity → see later session on Evaluating research
- Scales should be administered and scored according to the guidelines provided by the creators, and results should be interpreted in context and with caution (e.g. NASA-TLX is often administered with the wrong answer format)

System Usability Scale (SUS)

- Ten statements; respondents answer on a 1-5 scale whether or not they agree
- Generally used after the respondent has had an opportunity to use the system, but before any debriefing or discussion
- Respondents should record their immediate response to each item, rather than thinking about items for a long time
- All items should be checked. If a respondent feels that they cannot respond to a particular item, they should mark the centre point of the scale

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
2. I found the system unnecessarily complex	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
3. I thought the system was easy to use	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
5. I found the various functions in this system were well integrated	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
6. I thought there was too much inconsistency in this system	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
8. I found the system very cumbersome to use	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
9. I felt very confident using the system	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.

System Usability Scale (SUS)

- SUS yields a single number representing the **overall usability** of the system
- To calculate the SUS score:
 - Sum the score contributions from each item.
 - Each item's score contribution will range from 0 to 4
 - For items 1,3,5,7,and 9 the score contribution is the scale position minus 1
 - For items 2,4,6,8 and 10, the contribution is 5 minus the scale position
- Multiply the sum of the scores by 2.5 to obtain the overall value of SU
- The result is a score between 0 and 100

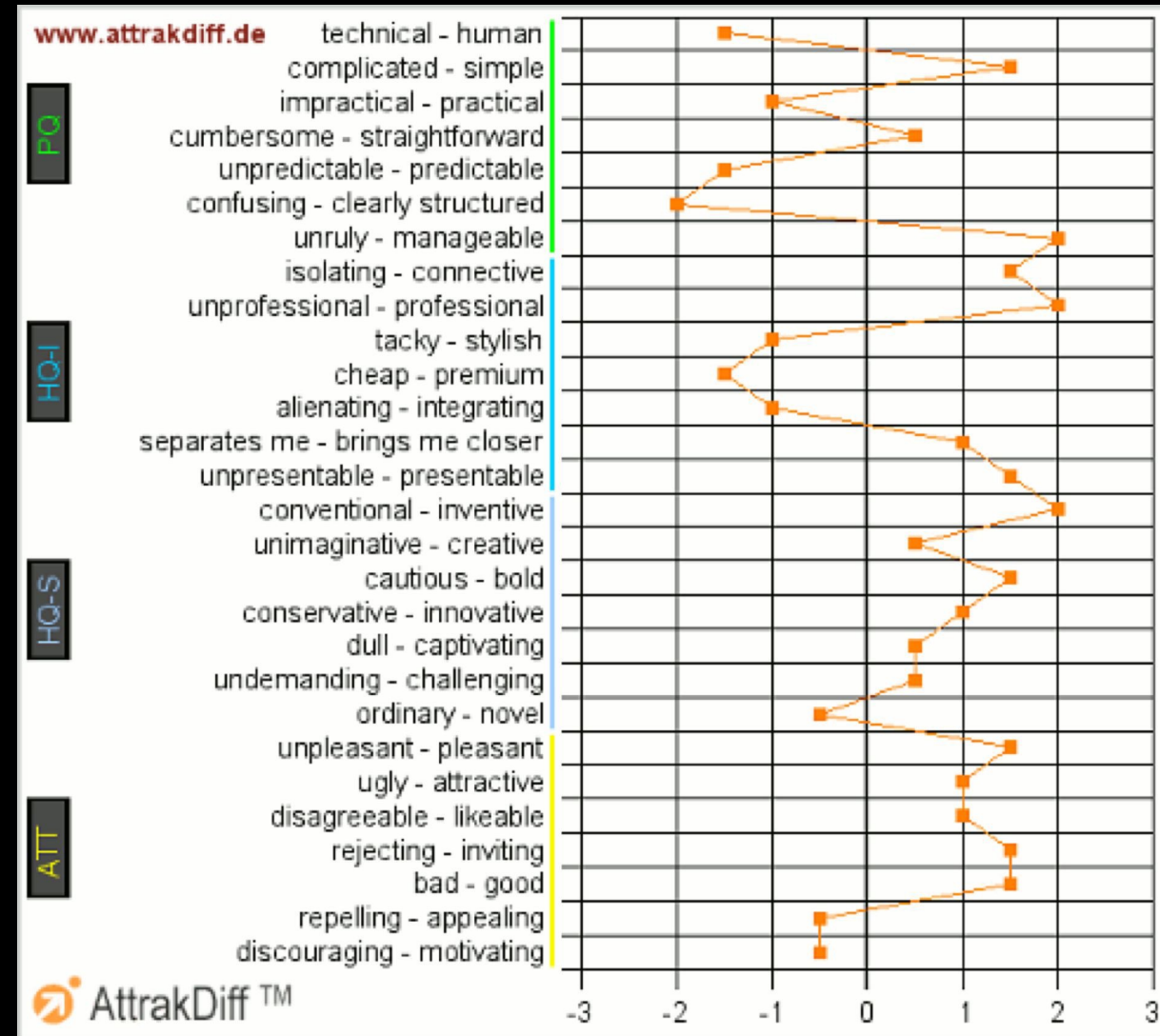
	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
2. I found the system unnecessarily complex	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
3. I thought the system was easy to use	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
5. I found the various functions in this system were well integrated	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
6. I thought there was too much inconsistency in this system	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
8. I found the system very cumbersome to use	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
9. I felt very confident using the system	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	1	2	3	4	5

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.

Attrakdiff

User Experience

- Measures subjective assessments concerning pragmatic and hedonic qualities and the attractiveness of interactive products
- 28 seven-step items whose poles are opposite adjectives (e.g. "confusing - clear", "unusual - ordinary", "good - bad").



Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003* (pp. 187-196). Vieweg+ Teubner Verlag.

AttrakDiff

hedonic

adjective • formal

UK  /hi:'dɒn.ɪk/ US  /hi:'dɑː.nɪk/

connected with feelings of pleasure:

- *Many purchases are related to hedonic impulses.*

pragmatic

adjective

UK  /præg'mæt.ɪk/ US  /præg'mæt.ɪk/



C2

solving problems in a sensible way that suits the conditions that really exist now, rather than obeying fixed theories, ideas, or rules:

- *In business, the pragmatic **approach** to problems is often more successful than an idealistic one.*

AttrakDiff

Norms in questionnaires

User Experience

- Measures subjective assessments concerning pragmatic and hedonic qualities and the attractiveness of interactive products
- 28 seven-step items whose poles are opposite adjectives (e.g. "confusing - clear", "unusual - ordinary", "good - bad")
- Alternatively use the User experience questionnaire UEQ



Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003* (pp. 187-196). Vieweg+ Teubner Verlag.

NASA-TLX





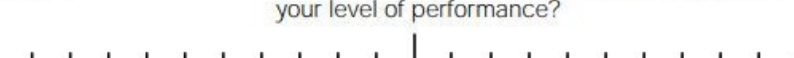

Assesses work load in six categories:

- Mental demand
- Physical demand
- Temporal demand
- Performance
- Effort
- Frustration

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date
Mental Demand	How mentally demanding was the task?	
		
Very Low		Very High
Physical Demand	How physically demanding was the task?	
		
Very Low		Very High
Temporal Demand	How hurried or rushed was the pace of the task?	
		
Very Low		Very High
Performance	How successful were you in accomplishing what you were asked to do?	
		
Perfect		Failure
Effort	How hard did you have to work to accomplish your level of performance?	
		
Very Low		Very High
Frustration	How insecure, discouraged, irritated, stressed, and annoyed were you?	
		
Very Low		Very High

Checklist for administering questionnaires

- Always use the original scales and only adapt items if necessary → items may need to be validated again
- Use clear and concise language in the questionnaire. Avoid using jargon or complex phrases that may be difficult for participants to understand.
- Avoid leading or biased questions. A leading question is one that suggests a particular answer, while a biased question is one that is framed in a way that could influence the participant's response.
- Use a variety of question types. This can help to ensure that you are collecting a range of data and that the questionnaire is not too repetitive.
- Test the questionnaire on a small group of participants before administering it to the full sample. This can help to identify any issues with the questionnaire and allow you to make any necessary revisions.
- Ensure that the questionnaire is appropriate for the age and abilities of the participants. For example, if you are administering the questionnaire to children, you may need to use simpler language or visual aids to help them understand the questions.

Spheres-of-Control Battery Items

Personal Efficacy

Sphere-specific measures of perceived control.

[D Paulhus](#) - Journal of personality and social psychology, 1983 - psycnet.apa.org

Proposes that individual differences in perceived control be partitioned into components associated with 3 primary spheres of behavior:(a) personal efficacy (control over the nonsocial environment as in personal achievement),(b) interpersonal control (control over other people in dyads and groups), and (c) sociopolitical control (control over social and political events and institutions). Assessment instruments are presented for measuring perceived control in each of these 3 spheres. Using data from 87 undergraduates, a 3-factor ...

☆ Save 📄 Cite Cited by 918 Related articles All 7 versions Web of Science: 367

[PDF] researchgate.net

	Disagree						Agree
When I get what I want it's usually because I worked hard for it.							
When I make plans I am almost certain to make them work.							
I prefer games involving some luck over games requiring pure skill							
I can learn almost anything if I set my mind to it.							
My major accomplishments are entirely due to my hard work and ability.							
I usually don't set goals because I have a hard time following through on them							
Competition discourages excellence.							
Often people get ahead just by being lucky							
On any sort of exam or competition I like to know how well I do relative to everyone else.							
It's pointless to keep working on something that's too difficult for me.						05.01.24	19

Checklist for formulating items

Comprehensability and Unambiguity are key to formulating items

- Formulate items positively and avoid negations; avoid double negations
- Avoid too complicated sentence (e.g. no nesting)
- No abbreviations or technical terms
- Avoid universal expressions such as "always", "never", "all"
- Define difficult concepts if necessary
- Pay attention to the respondent's prior knowledge and the target group
- Clearly specify time periods
- Avoid asking for intensities or frequencies

In-class exercise: Read through the Personal efficacy scale and identify problems with the items (5 minutes)

Custom Questionnaires

- Very commonly used in user research
- Custom questionnaires are useful if there are particular aspects to evaluate or particular information to gain
- Often in the form of Likert statements (agree-disagree), but can be open-ended questions, multiple choices, ordering tasks, etc.

5. Please state how comfortable you would be INTERACTING with a system (a large display) in the following locations *

1 = uncomfortable, 7 = comfortable

	1	2	3	4	5	6	7
Library	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Office workspace	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City center	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Cultural center	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Swimming hall (lobby)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
University campus	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shopping mall	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bus stop	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Checklist for administering questionnaires

- Always use the original scales and only adapt items if necessary → items may need to be validated again
- Use clear and concise language in the questionnaire. Avoid using jargon or complex phrases that may be difficult for participants to understand.
- Avoid leading or biased questions. A leading question is one that suggests a particular answer, while a biased question is one that is framed in a way that could influence the participant's response.
- Use a variety of question types. This can help to ensure that you are collecting a range of data and that the questionnaire is not too repetitive.
- Test the questionnaire on a small group of participants before administering it to the full sample. This can help to identify any issues with the questionnaire and allow you to make any necessary revisions.
- Ensure that the questionnaire is appropriate for the age and abilities of the participants. For example, if you are administering the questionnaire to children, you may need to use simpler language or visual aids to help them understand the questions.

Subjective measures

Pro's

- Easy to
 - administer
 - Analyze
- Constructs are well-defined
- Inexpensive
- Fast

Problems

- Introspection is necessary
- Can often only be gathered retrospectively; after things happened
- May not be appropriate for every
 - Language
 - Culture
- Prone to
 - Careless responding
 - Cognitive biases (e.g., response bias)
 - Deception
- Standardisation and Validation is rare in some fields (e.g. HCI)

Behavioral measures

- Anything that measures observable behavior
- In engineering psychology these assess how people perform tasks, make decisions, and interact with technology



**Imagine you are an XRAY-
Technician in Vantaa**

Please make a list from 1 to 6
6 X-rays of suitcases are briefly
presented. Your task is to detect if
there is a weapon inside.

Note down whether there was a
gun or not.

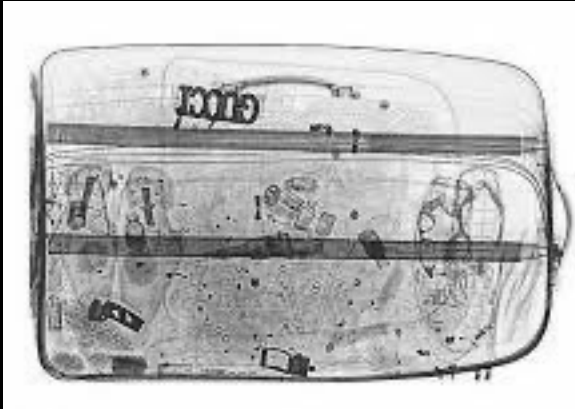














SOLUTION

Was there a gun?

1. NO

2. YES

3. NO

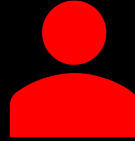
4. NO

5. NO

6. YES

SOLUTION

Was there a gun?



1. NO

2. YES

3. NO

4. NO

5. NO

6. YES

1. NO

2. YES

3. NO

4. YES

5. NO

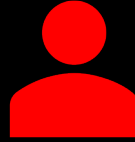
6. YES

False alarm

SOLUTION

Was there a gun?

- 1. NO
- 2. YES
- 3. NO
- 4. NO
- 5. NO
- 6. YES



- 1. NO
- 2. YES
- 3. NO
- 4. YES
- 5. NO
- 6. YES

← False alarm



- 1. NO
- 2. NO
- 3. NO
- 4. NO
- 5. NO
- 6. YES

← MISS

Accuracy

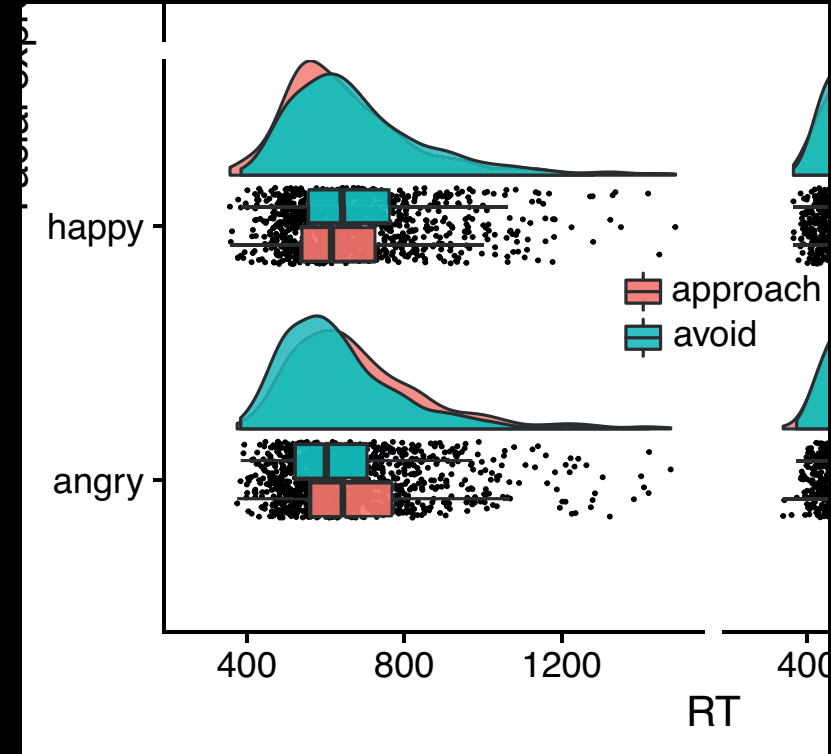
- A measure how accurately a person performs a task or responds to a stimulus
- Affected by similar cognitive factors as response time
- Accuracy can be measured as:
 - Percentage of correct responses: The proportion of correct responses made by a person during a task.
 - Absolute error: The difference between a person's response and the correct answer.
 - Percent error: The difference between a person's response and the correct answer, expressed as a percentage of the correct answer
- Does not need to be binary, e.g. lane displacement



	Noise	Signal
Yes	False Alarm	Hit
No	Correct Rejection	Miss

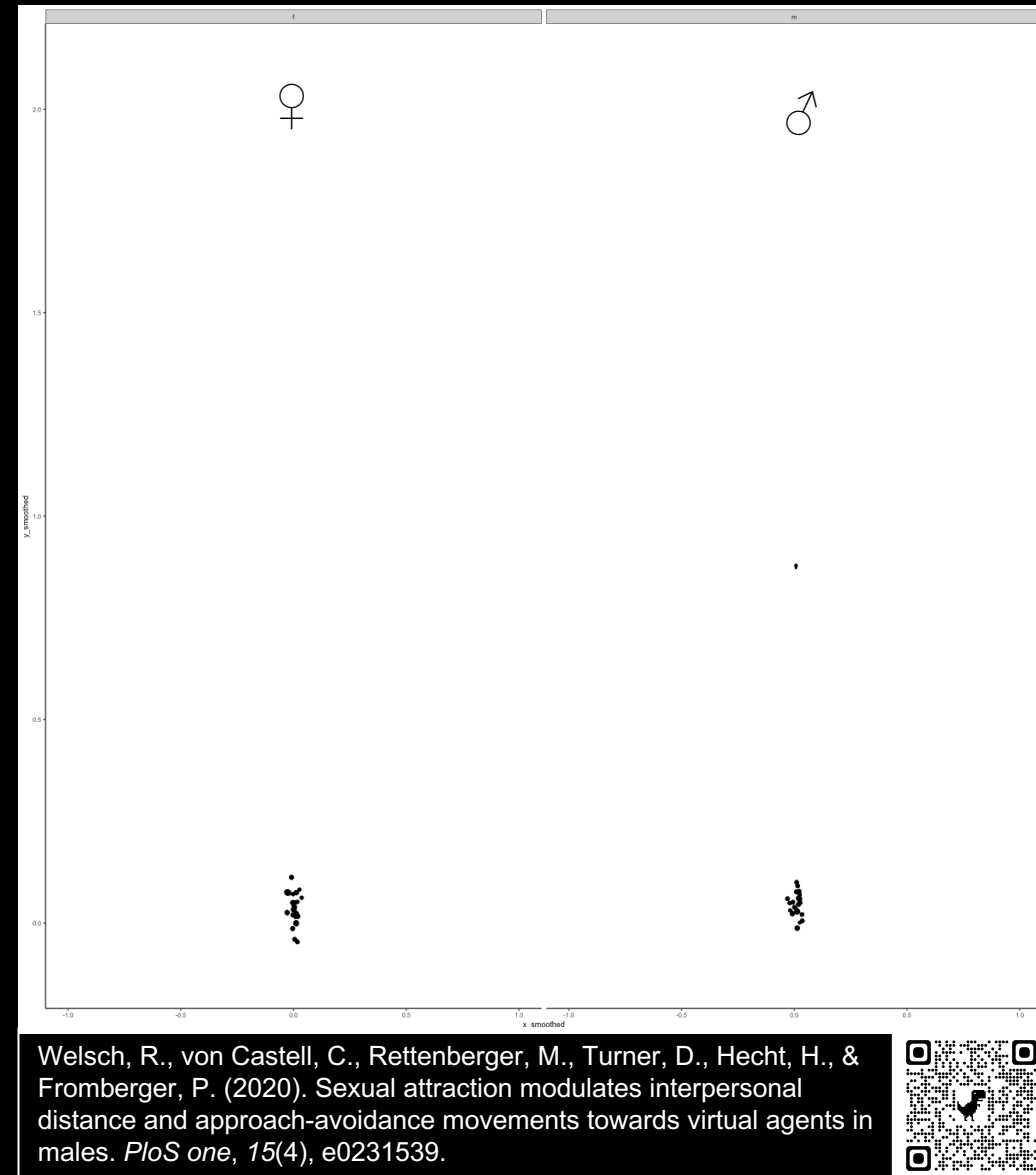
Response times

- Response time is a measure of how quickly a person responds to a stimulus or performs a task
- Response time can be affected by a variety of factors, including
 - the complexity of the task
 - the availability of cognitive resources (e.g., attention, memory)
 - Personality (also person's skill and experience)
 - Competing tasks
- Data is often non-normal
- Often measured for choices, simple responses, discrimination of stimuli, ratings, or task completion



Movement data

- Information on objects or a person in space and time
- can be collected with:
 - Motion capture systems: These use sensors or markers placed on the body to track movements in three-dimensional space.
 - Wearable sensors: These use sensors such as accelerometers or gyroscopes to measure movement in one or more dimensions.
 - Video tracking: This involves analyzing video footage of a person's movements to extract kinematic data
 - ...
- Can inform on psychological processes such as motor control, coordination, and attention
- Movement data is often used in engineering psychology and human factors research to study how people interact with technology, to design ergonomic environments, and to evaluate the effectiveness of training programs




Direct vs. indirect measures

Viewing time in forensic psychology

- Overt measurement of one variable while covert measurement of other variables
- Ethical problem of deceiving the participant
- Overt measurement can influence the indirect measurements

How sexually attractive do you find this person?



1 2 3 4 5 6 7
Very unattractive Neutral Very attractive

NEXT

Pezzoli, P., Babchishin, K., Pullman, L., & Seto, M. C. (2022). Viewing time measures of sexual interest and sexual offending propensity: an online survey of fathers. *Archives of sexual behavior*, 51(8), 4097-4110

Welsch, R., Schmidt, A. F., Turner, D., & Rettenberger, M. (2021). Test-retest reliability and temporal agreement of direct and indirect sexual interest measures. *Sexual Abuse*, 33(3), 339-360.

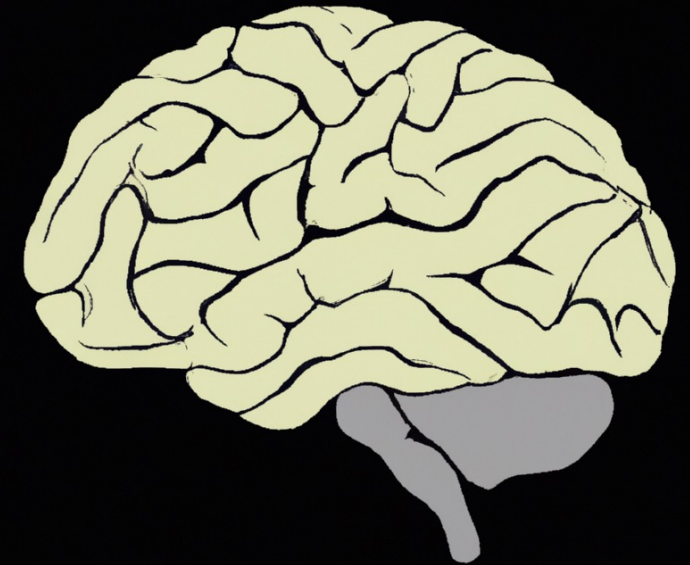


Checklist behavioral measures

- Is the task appropriately defined?
- How objective is the measurement?
- Are all behavioral events considered?
- Does the measured behavior map onto the psychological construct?
- Does measurement affect behavior?
- Is the context controlled and/or recorded?

Psychophysiological measures

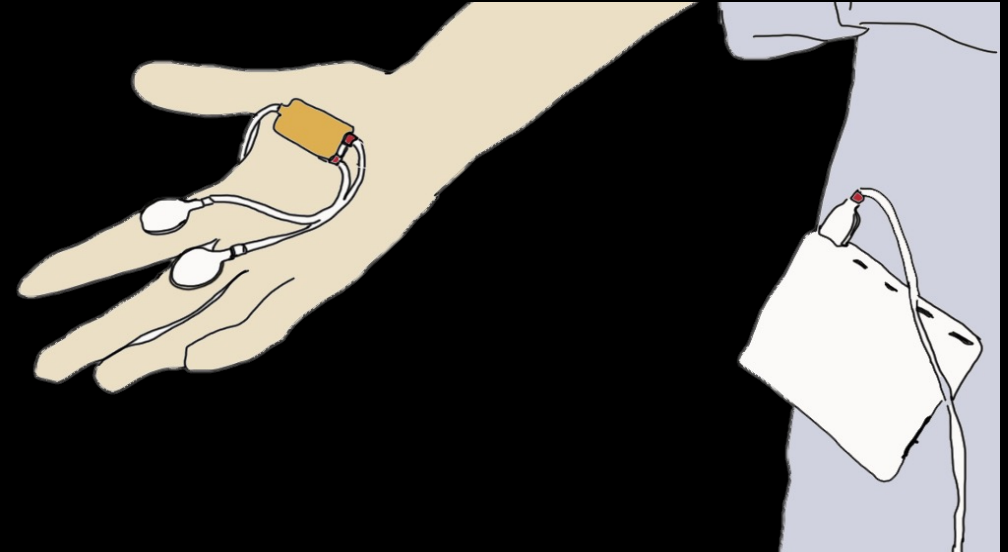
- Psychophysiology refers to the study of the physiological processes that underlie psychological phenomena, such as emotions, cognition, and behavior.
- Psychophysiological measures are techniques used to assess physiological activity, such as heart rate, skin conductance, and brain activity.
- These measures can provide valuable insights into psychological processes, as they offer a direct and objective measure of physiological responses.
- Common psychophysiological measures include electroencephalography (EEG), electromyography (EMG), and electrocardiography (ECG).
 - They can often be measures online (as the participant thinks, feels or experiences)
 - They suffer from noise and are sometimes unspecific



DEMO-Time

EDA

- Electrodermal activity (EDA) refers to changes in the electrical conductance of the skin
- reflect changes in sweat gland activity, which is influenced by the sympathetic nervous system
- is often used as a measure of emotional arousal, as it is thought to increase in response to emotionally significant events or stimuli
- sensitivity to factors such as temperature and humidity, as well as the potential for habituation to occur over repeated testing
- Two components
 - Phasic arousal refers to rapid, transient changes in EDA that are typically associated with emotional or cognitive processing
 - Tonic arousal refers to sustained or baseline levels of EDA that are present even in the absence of specific stimuli.

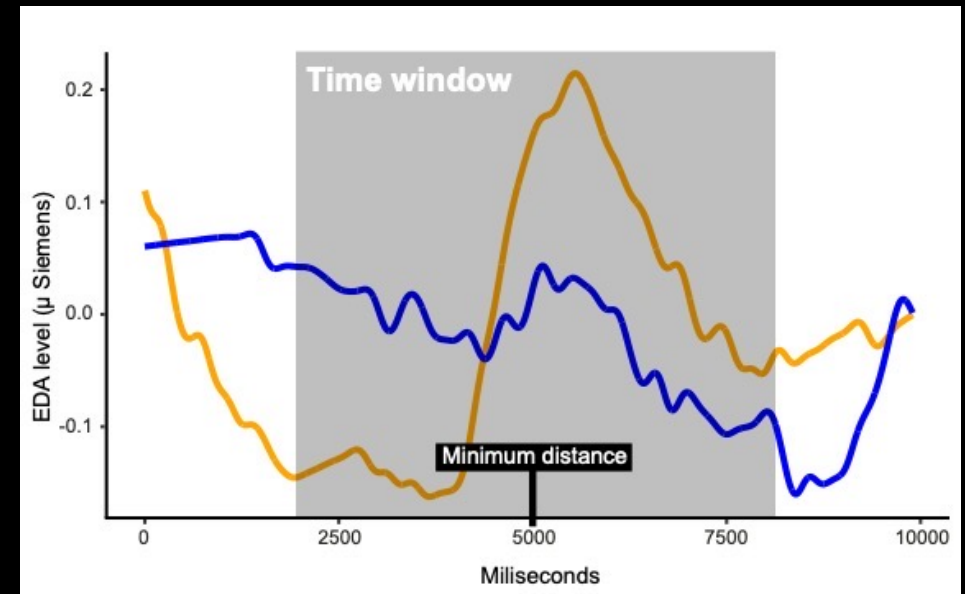


Huang, A., Knierim, P., Chiossi, F., Chuang, L. L., & Welsch, R. (2022, April). Proxemics for Human-Agent Interaction in Augmented Reality. In *CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Evoked / Event-related potentials

are often measured using electroencephalography (EEG), electromyography (EMG), or other techniques, such as magnetoencephalography (MEG) or positron emission tomography (PET)

- Measurements are time-locked to specific events or stimuli
- differ from continuous measures in that they only reflect brain activity or other physiological responses in response to specific stimuli or events, rather than ongoing activity
- sensitive to subtle changes in cognitive or physiological processes
- Can be contrasted between stimuli/conditions



Huang, A., Knierim, P., Chiossi, F., Chuang, L. L., & Welsch, R. (2022, April). Proxemics for Human-Agent Interaction in Augmented Reality. In *CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Invasive vs. non-invasive measures

- invasive measures involve the insertion of electrodes or other sensors into the body, often requiring surgery or other medical procedures
 - intracranial electrodes, deep brain stimulation, and single unit recording
 - Rarely done with human subjects
- Non-invasive measures do not involve the insertion of electrodes or other sensors into the body
 - can often be performed without the need for specialized medical training
 - electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and transcranial magnetic stimulation (TMS)
- Psychology typically uses non-invasive measures due to resources and ethical considerations

Eye-Tracking

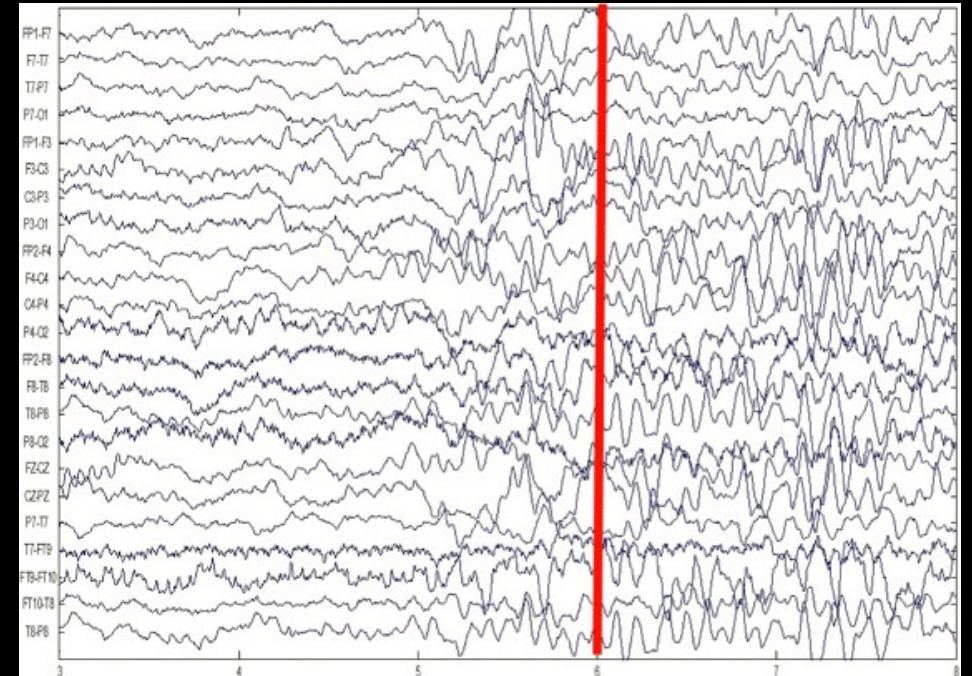
- measurement of eye movements and gaze patterns in response to visual stimuli
- Is used to study attention, perception, memory, and cognitive processing
- Indices
 - Fixation duration: the amount of time spent looking at a specific location or stimulus
 - Saccade amplitude: the distance between two consecutive fixations
 - Gaze direction: the direction of the eye gaze in relation to the visual field
 - Scanpath: the sequence of fixations and saccades made while viewing a visual stimulus
 - Pupil size: the size of the pupil, which can reflect arousal or cognitive effort
- In HCI: Dwell time for areas of interest for an interface can be measured



https://www.tiktok.com/@freighttrainofficial/video/7187485107610242350?_r=1&_t=8YyUHxPtG50

Electroencephalography (EEG)

- EEG is a technique that involves the measurement of electrical activity in the brain using electrodes placed on the scalp.
- Oscillations
 - Alpha power: the amount of electrical activity in the alpha frequency range (8-12 Hz), which is thought to reflect relaxation and attention
 - Beta power: the amount of electrical activity in the beta frequency range (13-30 Hz), which is thought to reflect cognitive effort and arousal
 - Theta power: the amount of electrical activity in the theta frequency range (4-7 Hz), which is thought to reflect memory and learning
- Event-related potentials
 - Time-locked amplitudes that follow a certain pattern
 - E.g. N400 (negative polarization after word that is not understood)



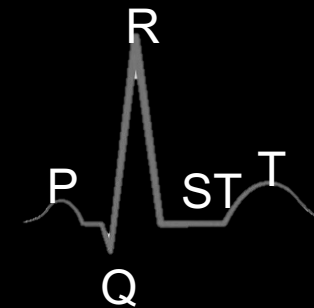
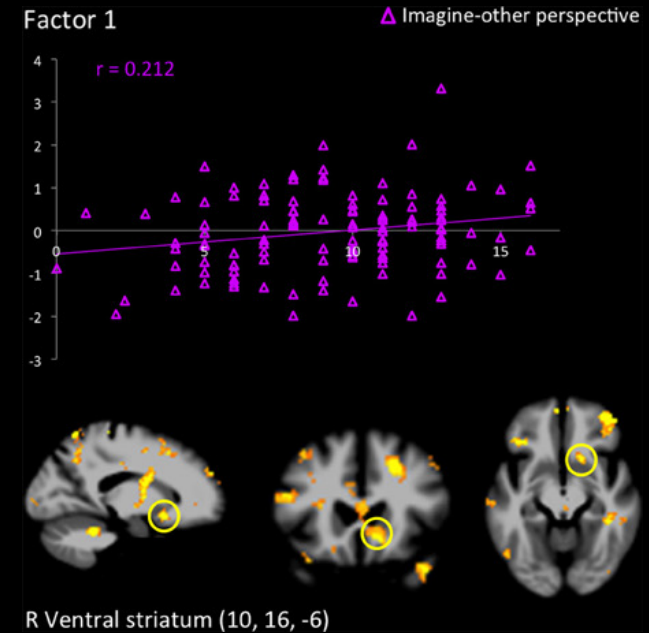
Xun, G., Jia, X., & Zhang, A. (2016). Detecting epileptic seizures with electroencephalogram via a context-learning model. *BMC medical informatics and decision making*, 16(2), 97-109.

Other common measures

- fMRI: Blood oxygen-level in the brain
- fNIRS: Bloodoxygenisation measured by changes of reflected light (infared) on the scalp
- ECG: electrocardiogram measures heart's electrical activity
-

<https://www.aclsmedicaltraining.com/basics-of-ecg/>

Decety, J., Chen, C., Harenski, C., & Kiehl, K. A. (2013). An fMRI study of affective perspective taking in individuals with psychopathy: imagining another in pain does not evoke empathy. *Frontiers in human neuroscience*, 7, 489.



Checklist for psychophysiological data

- Is all equipment working and properly calibrated?
- Were participants in a neutral state before starting the test?
- Have you done a baseline measurement?
- Are any artifacts in the data (e.g. power line noise)?
- Was the timing synched appropriately (time-stampes, markers)?
- Are any medical conditions not specific to the study affecting the recording?
- Is the signal-noise ratio appropriate?

**How do interpret
measurement?**

Stevens' theory of measurement

A conceptual framework for understanding the nature of measurement in psychology and the social sciences

- measurement involves the assignment of numerals to objects or events according to some rule or procedure, in order to represent the attributes or qualities of those objects or events.
- Four levels of measurement
 - Nominal
 - Ordinal
 - Interval
 - Ratio

Stevens' theory of measurement

A conceptual framework for understanding the nature of measurement in psychology and the social sciences

Nominal: This involves the assignment of labels or categories to objects or events, without any inherent order or magnitude.

- For example
 - Likes pineapple on pizza (yes vs. No)
 - Ethnicity
 - Family status
 - Blood-type
 -



Stevens' theory of measurement

A conceptual framework for understanding the nature of measurement in psychology and the social sciences

Ordinal: This involves the assignment of rankings or orderings to objects or events, but not necessarily equal intervals or distances between them.

- For example
 - Grades
 - Tax groups
 - Likert-scales or preference ratings (e.g., very satisfied, satisfied, dissatisfied)
 -

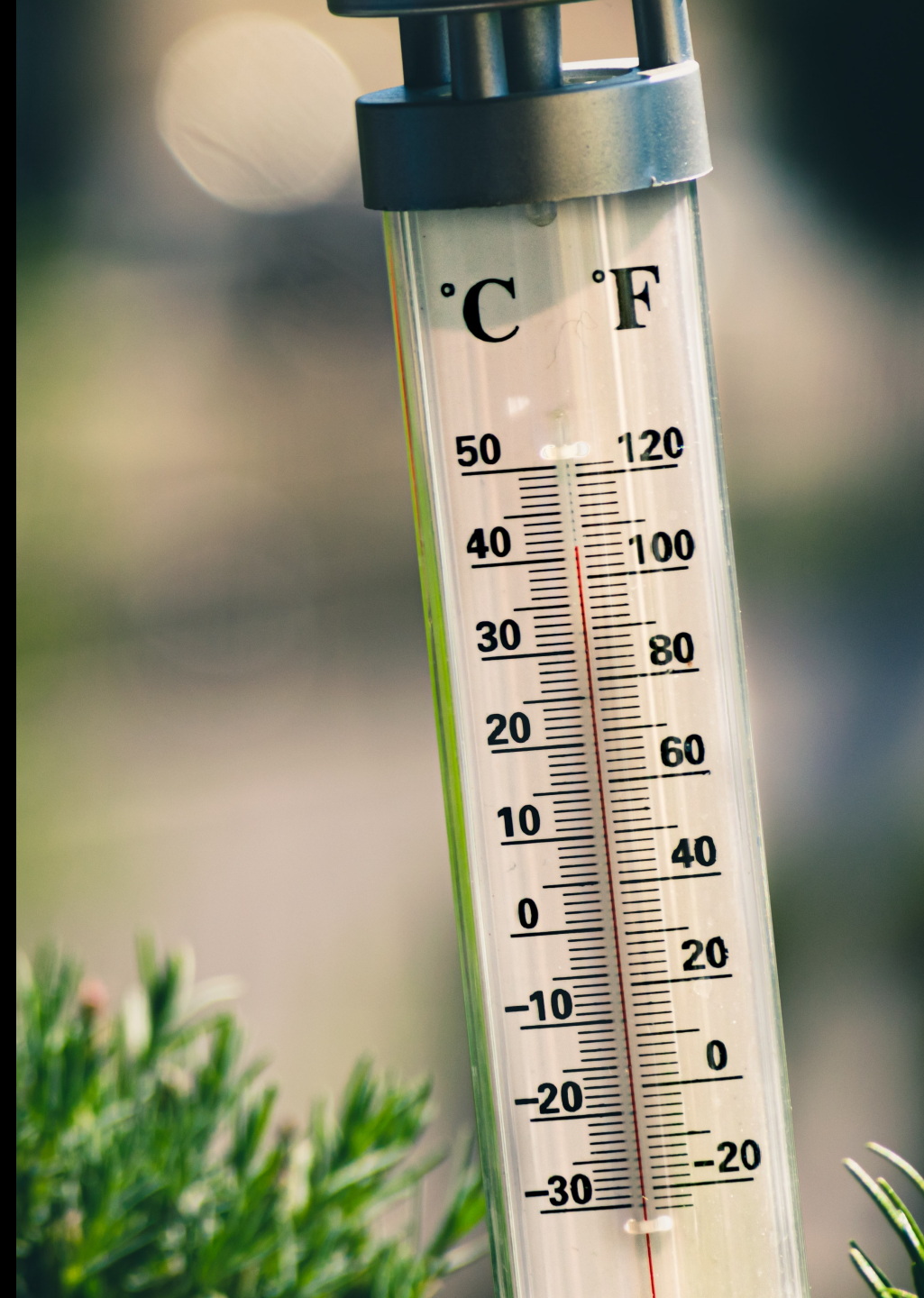


Stevens' theory of measurement

A conceptual framework for understanding the nature of measurement in psychology and the social sciences

Interval: This involves the assignment of numbers to objects or events such that equal intervals between numbers correspond to equal differences in the attribute being measured.

- For example
 - Thermometer
 - IQ (take an IQ-test here: <https://openpsychometrics.org/tests/FSTIQ/>)
 - Psychopathy
 - BIG-Five
 -



Stevens' theory of measurement

A conceptual framework for understanding the nature of measurement in psychology and the social sciences

Ratio: This involves the assignment of numbers to objects or events such that there is a clear and meaningful zero point, and equal intervals between numbers correspond to equal ratios of the attribute being measured.

- For example
 - Height
 - Weight
 - Speed
 - Area
 - Distance
 -



Stevens' theory of measurement

A conceptual framework for understanding the nature of measurement in psychology and the social sciences

- measurement involves the assignment of numerals to objects or events according to some rule or procedure, in order to represent the attributes or qualities of those objects or events.

- Four levels of measurement

In Psychology numeric variables are often treated as interval-scaled, while in HCI they are rather deemed to be ordinal

- Nominal: This involves the assignment of labels or categories to objects or events without any inherent order or magnitude.

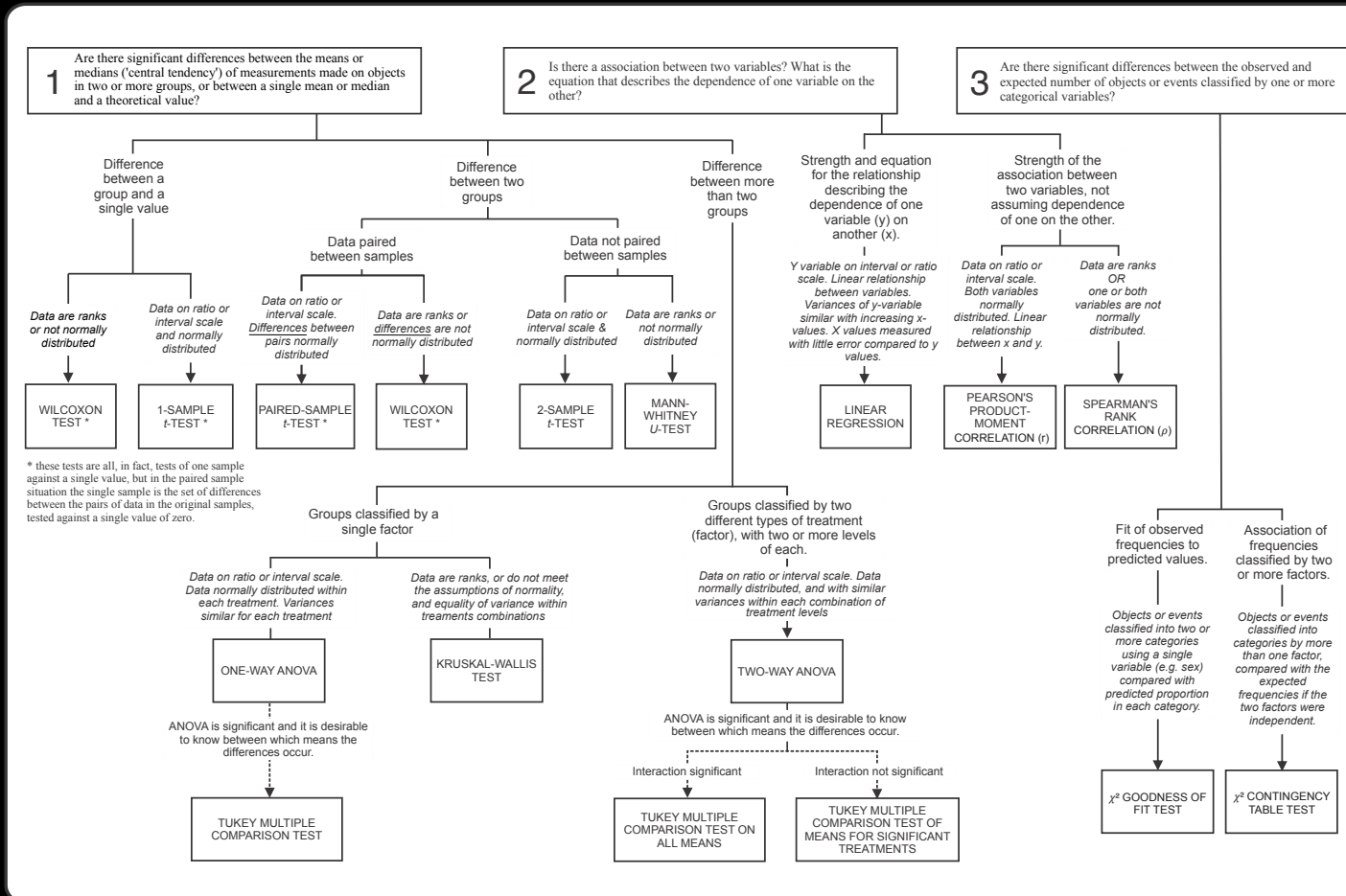
- Ordinal: This involves the assignment of rankings or orderings to objects or events without necessarily involving any distance between them.

- Interval: This involves the assignment of numbers to objects or events such that equal intervals between numbers correspond to equal differences in the attribute being measured.

- Ratio: This involves the assignment of numbers to objects or events such that there is a clear and meaningful zero point, and equal intervals between numbers correspond to equal ratios of the attribute being measured.

Choosing a statistical model

Hypothesis x Study design x Measurement scale = Statistical model



Note:

- All statistical models are wrong as they are only approximations of the real world
- Different disciplines use different models
 - Economics: Regression
 - Psychology: ANOVA
 - They are essentially the same but differ in their reporting

Only for NHST stats

Summary

Psychological measurement

- Psychological measurement often relates to immeasurable constructs
- We need to operationalize constructs of interest
- Measurement can be done on a subjective, physiological or behavioral level
- Data can typically be categorized into four levels according to Steven's theory of measurement which has an impact on the use of data

Questions?

Evaluating empirical research

- Scientific Discourse
- Sampling
- Objectivity
- Reliability
- Validity







Scientific Discourse

essential for advancing our understanding of the world and making informed decisions based on evidence

- Scientific discourse helps to advance knowledge and understanding within a field by encouraging critical thinking and debate
- It allows for the testing and refinement of hypotheses and theories, and helps to identify gaps in current knowledge
- Scientific discourse helps to ensure that research is conducted in an ethical and transparent manner, and helps to prevent fraud or bias in the scientific process
- Evaluation of scientific studies involves the careful examination of the methods, results, and conclusions of a study to determine its reliability and validity
- ensure that research is of high quality and can be trusted

Discourse about empirical studies

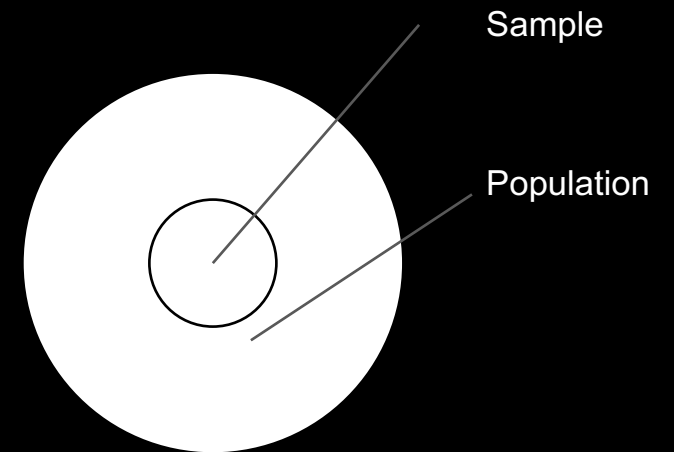


WEIRD Samples

- WEIRD samples are those that are primarily drawn from **Western, educated, industrialized, rich, and democratic** societies, such as the United States and Western Europe.
- can lead to overgeneralization
- sampling bias, as they may not be representative of the general population
 - may only reflect the experiences and perspectives of a specific group
- may perpetuate cultural stereotypes

Sampling

- refers to the process of selecting a subset of individuals/stimuli from a larger group, or population, to study or represent the population as a whole
- make inferences about the population based on the characteristics of the sample
- sample characteristics may affect the validity of the results → does the sample represent the population we talk about?
- Also items can be considered a sample of a population (typically we cannot sample all potential items)



Types of samples

Most common sampling methods

- Random sampling
 - selecting individuals from the population randomly, using a randomization method, to ensure that every individual has an equal chance of being selected
 - E.g. select every 10th name on a list
- Representative & Stratified sampling
 - divide the population into subgroups (strata) based on certain characteristics and then randomly selecting individuals from each stratum
 - E.g. divide the population of all college students into subgroups based on major (e.g. science, engineering, humanities) and select a sample from each subgroup
- Cluster sampling involves dividing the population into smaller groups (clusters) and selecting a sample of clusters to study
 - E.g. divide sample by postcode and sample from these
- Convenience sampling
 - selecting individuals who are easily accessible or willing to participate, which may not be representative of the population as a whole
 - E.g. ask students on campus to participate

How big should your sample be?

Five approaches

1. Power analysis

- The effect size is known and power (probability of rejecting H_0 when it is false) is specified
- Stop when sample size can reliably find **any** effect

2. Resource constraints

- Effect size is not relevant
- Stop when resources (e.g., money) are depleted

3. Smallest effect of interest

- The effect size is known and power (probability of rejecting H_0 when it is false) is specified
- Stop when sample size can reliably find this smallest effect of interest

4. Precision analysis

- The effect size is not known
- Stopping on a statistical estimation criterion (e.g., $CI_{upper} \geq 0$ & $CI_{lower} \geq 0$)

5. Sequential analysis

- The effect size is not known
- Stopping on a statistical inference criterion (e.g., $p < .00001$)

Inference

	H_0 not rejected	Rejection of H_0
Reality	H_0 is true	False negative False positive (alpha)
	H_0 is false	True negative True positive (power)

Decision Matrix in NHST

Psychometrics

A subfield in psychological science

1. Theory of measurement
2. Development of novel measurements (e.g., questionnaires)
3. Evaluation of measurements
 - Objectivity
 - Reliability
 - Validity

Objectivity

- Objectivity refers to the ability to observe and assess situations, events, or data without personal bias or prejudice
- Ensure that results are accurate
- Often quantified as the agreement between (multiple) raters
 - Measures how consistently different people or groups evaluate the same thing
 - high level of rater agreement indicates that the ratings are consistent and objective, while a low level of rater agreement may suggest bias or subjectivity
 - Statistics are Kappa coefficient (nominal data), correlation (ordinal, or interval data with two raters), and the intraclass correlation coefficient (multiple raters)
- Often important for observation but less important for computerized testing

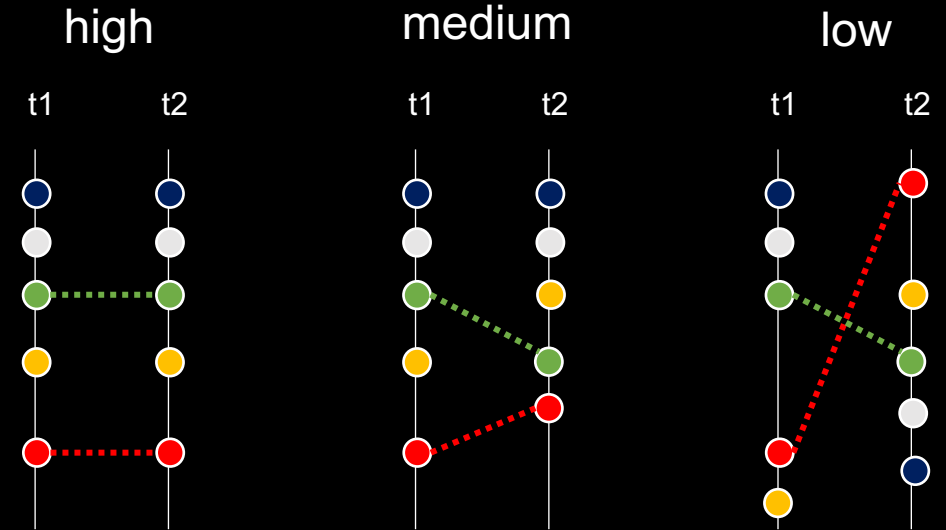
How reliable is my measure?

How much noise is in the signal? o is the measurement t is the true score

$$R = \frac{\sigma_t}{\sigma_o}$$

Reliability

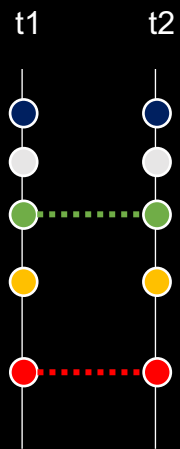
- Stability of a measure or instrument
- Test-retest reliability
 - stability of a measure over time, and can be assessed by administering the same/equivalent measure to the same group of people at two different points in time
- Internal reliability
 - Internal consistency
 - internal consistency reliability refers to the consistency of items within a measure or instrument
→ Cronbach's alpha (dividing the sum of the item variances by the total variance and subtracting the result from 1; lower bound of reliability)
 - Split-half reliability
 - dividing a test or measure into two halves and comparing the scores to determine the
 - consistency of the results → correlation



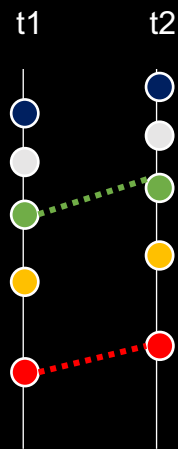
Problems with reliability

Always inspect your data!

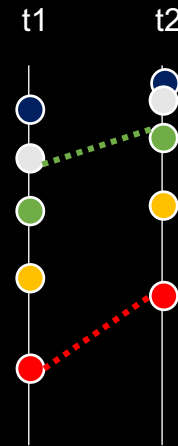
High reliability
Perfect agreement



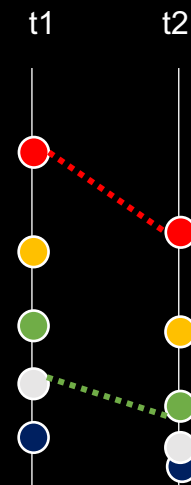
High reliability
low agreement



Ceiling effect



Floor effect



Validity

- Ability of a measure to accurately assesses what it is intended to measure
- Types of content validity:
 - Face validity
 - extent to which a test or measure appears to measure what it is intended to measure based on its content
 - Criterion validity
 - Predictive validity
 - Ability of the measure to accurately predict future outcomes or behavior
 - Concurrent validity
 - Relation to other measures that measure the same construct at the same time
- Construct validity:
 - Convergent validity
 - a measure correlates with other measures that are believed to assess the same construct or concept
 - Discriminant validity
 - a measure does not correlate with measures of unrelated constructs or concepts.
- Statistical validity: Are the statistical procedures appropriate to draw the conclusions
- Validity of study design

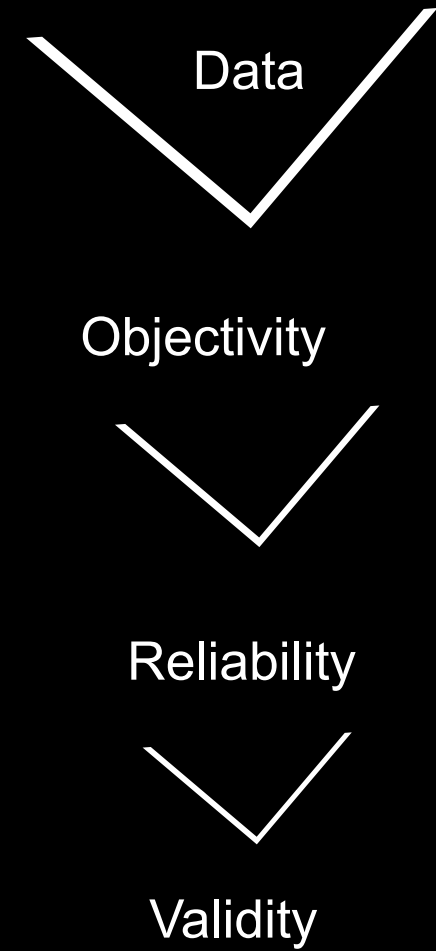
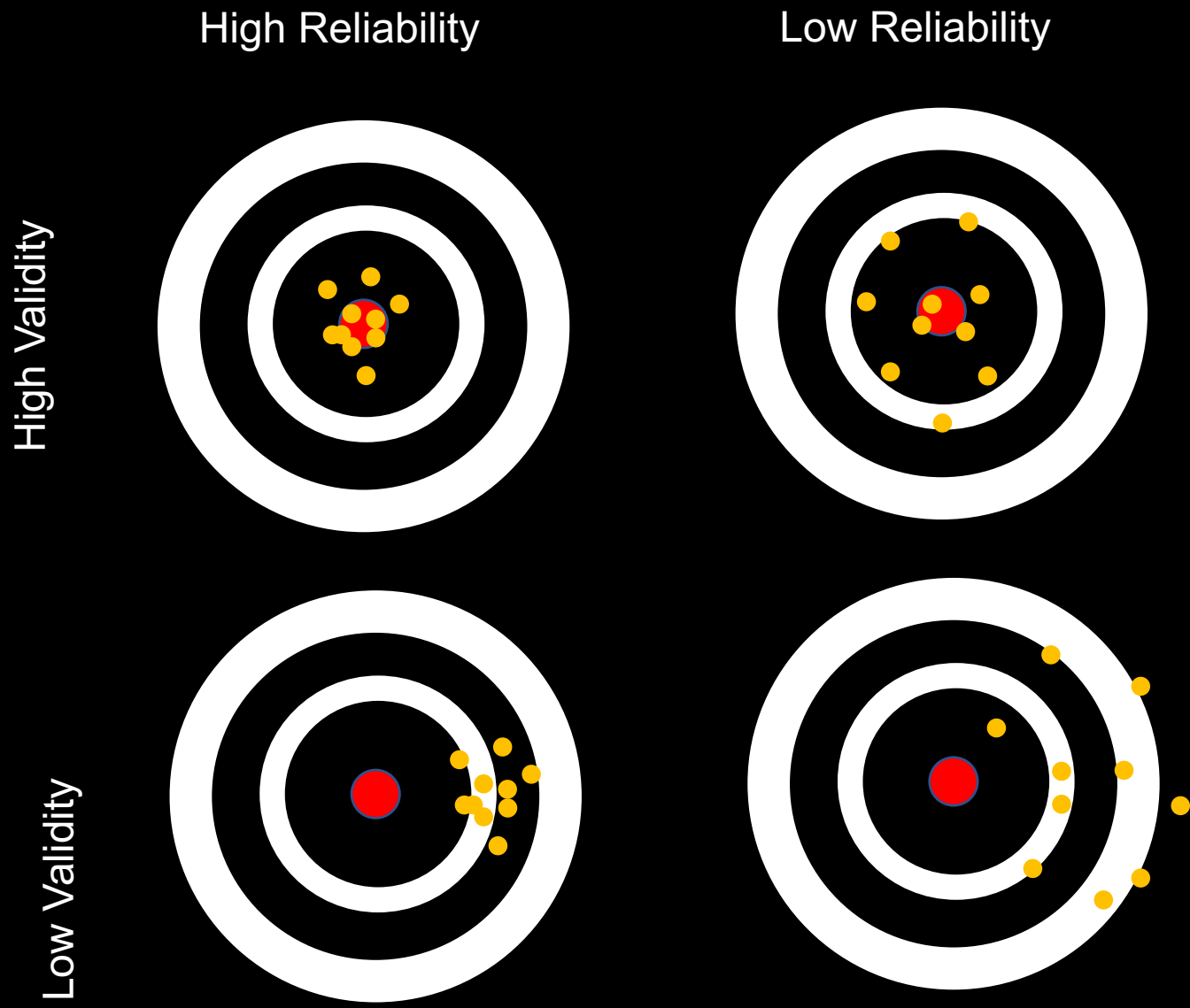
Validity in Experiments

Internal validity

- Counfounds are controlled
- Causality can be assumed
- Accuracy is prioritized

External validity

- Results can be applied to contexts outside the lab
- Resemble real-world contexts
- Generizability is prioritized



Summary

- Scientific discourse involves evaluating the quality and validity of research to determine its merit and usefulness
- Sampling is the process of selecting a subset of individuals or observations from a larger population in order to make inferences about the population as a whole.
- Objectivity involves minimizing bias and ensuring that research is conducted and reported in an unbiased manner.
- Reliability refers to the consistency and stability of research findings, which can be assessed through various methods.
- Validity refers to the extent to which a test or measure accurately assesses what it is intended to measure

You now have all the tools to plan your study!



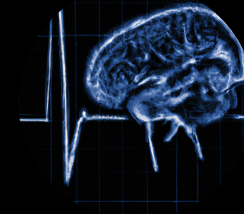
Psychological
Research



Advancing theory



Study designs and
variables



Measurement
in Psychology



Evaluating
empirical
research

Self-assessment

Lecture 3

1. Stevens theory of measurement: What level of measurement is this?
 - Consider the following variables: gender, IQ, political party affiliation, income level, grades, height and age.
 - For each variable, determine which level of measurement (nominal, ordinal, interval, or ratio) is most appropriate and state why
2. Skim through “Initial validation of the general attitudes towards Artificial Intelligence Scale”
(<https://doi.org/10.1016/j.chbr.2020.100014>)

What psychometric evaluation did they use?