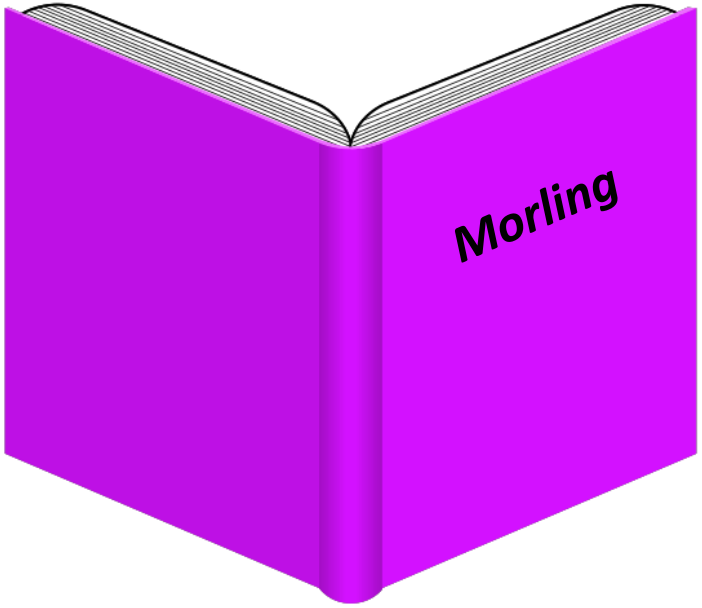


Research Methods and Statistics

Lecture 2: Measurement in Psychology

Riet van Bork

Pictures source: pixabay.org

Some claims about psychology

If you want someone to like you more, ask for their help.

Playing with blocks helps younger children with learning because it teaches spatial ability.

The color blue is an appetite suppressant.

Thinking through decisions in a foreign language helps to lower emotions and helps you focus on making a more rational decision.

Your core personality traits don't change as you age. However, over time your anxiety levels, friendliness and bravery for trying new things can flux multiple times depending on experiences and trauma related to those experiences.

Some claims about psychology

If you want someone to like you more, ask for their help.

Playing with blocks helps younger children with learning because it teaches spatial ability.

The color blue is an appetite suppressant.

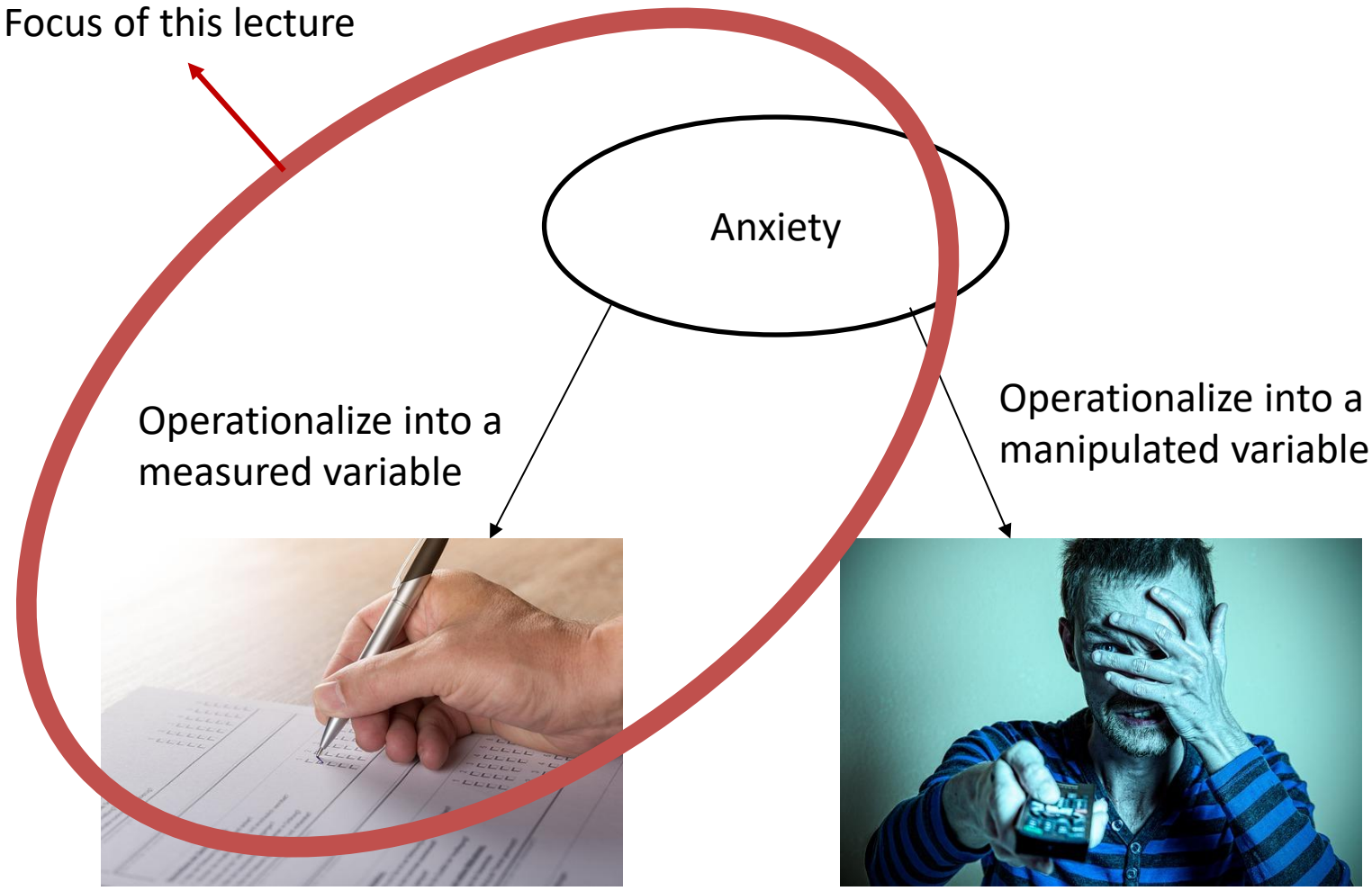
Thinking through decisions in a foreign language helps to lower emotions and helps you focus on making a more rational decision.

Your core personality traits don't change as you age. However, over time your anxiety levels, friendliness and bravery for trying new things can flux multiple times depending on experiences and trauma related to those experiences.

Psychological attributes

- All of these are *variables*: they *vary* (have different levels) across people or time (or both):
  - more vs less liking
  - better vs worse spatial ability
  - more vs less appetite
  - more vs less emotions and focus
  - higher vs lower levels of anxiety, friendliness and bravery
- Just like weight, length and volume, these are properties (often called ‘attributes’), but *psychological* properties (properties of the mind)
- To study these variables in empirical research, we have to operationalize these properties (turn them into something observable)
- For example, suppose you are interested in the relationship between anxiety and study performance..

Operationalizing anxiety



Pictures from pixabay.com

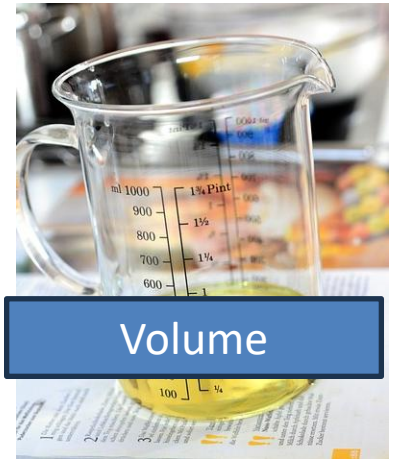
Today

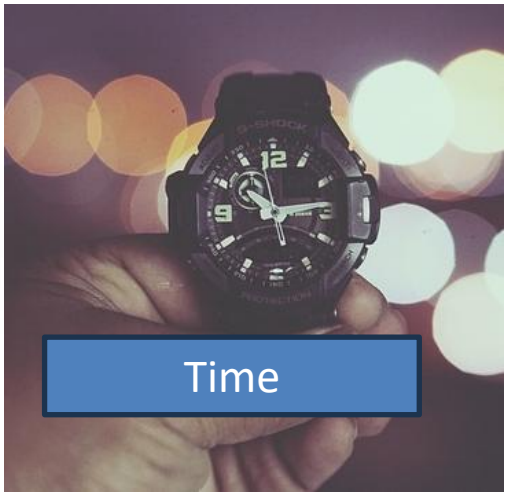
1. Measurement instruments in psychology
2. Problems
3. Quality of measurement in psychology

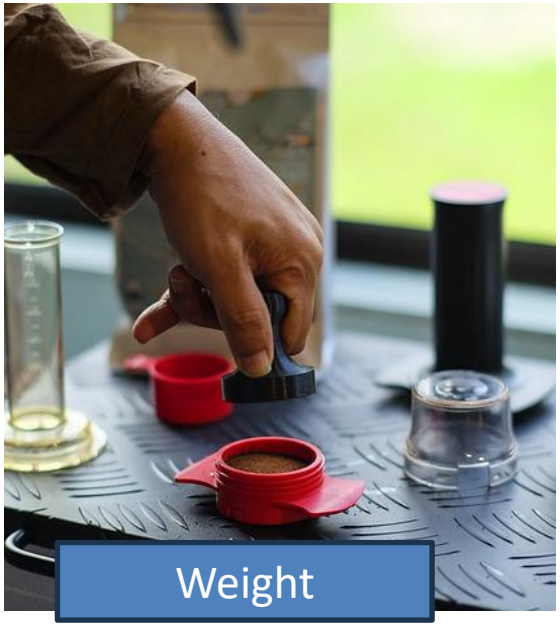
Measurement in daily life

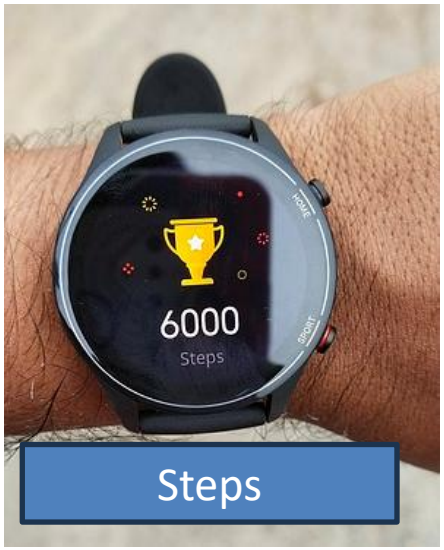
Pictures from pixabay.com

Measurement in daily life



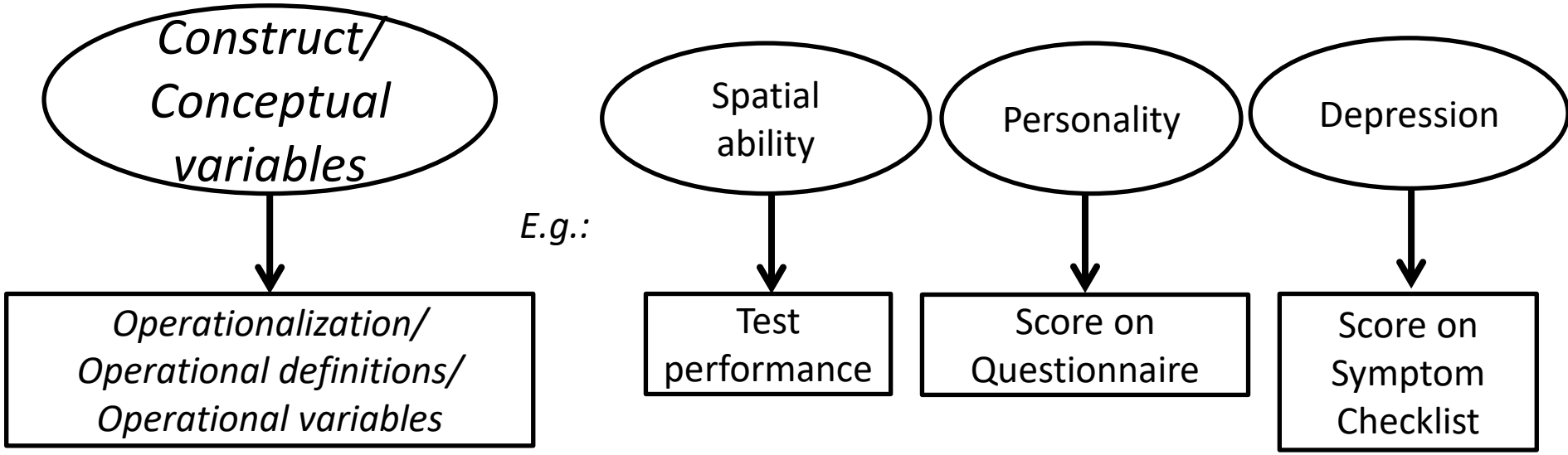






Pictures from pixabay.com

In Psychology



- Construct: Psychological attribute which is not directly observable
  - E.g.,: cognitive abilities, personality traits, opinions, emotions, mental disorders
  - Note: in articles, typically circles represent 'latent variables' and squares 'observed variables'

Personality

- Popular personality questionnaire: ‘Big 5’
  - Constructs:
    - Openness to experience
      - Tendency to prefer novelty and variety
    - Conscientiousness
      - A tendency to show self-discipline
    - Extraversion
      - The tendency to seek stimulation in the company of others
    - Agreeableness
      - The tendency to be cooperative and empathetic
    - Neuroticism
      - The tendendy to experience unpleasant emotions easily

Factor	Label	Adjective definers	NEO-PI-R facet scale definers
N	Neuroticism, Negative Affectivity vs. Emotional Stability	Calm—Worrying	Anxiety
		Even tempered—Temperamental	Angry hostility
			Depression
		Self-satisfied—Self-pitying	Self-consciousness
			Impulsiveness
		Comfortable—Self-Conscious	Vulnerability
		Unemotional—Emotional	
		Hardy—Vulnerable	
E	Extraversion, Surgency, Social Activity vs. Introversion	Reserved—Affectionate	Warmth
		Loner—Joiner	Gregariousness
		Quiet—Talkative	Assertiveness
		Passive—Active	Activity
		Sober—Fun loving	Excitement seeking
		Unfeeling—Passionate	Positive emotions
O	Openness to Experience, Intellect, Culture vs. Closedness	Down to earth—Imaginative	Fantasy
		Uncreative—Creative	Aesthetics
			Feelings
		Conventional—Original	Actions
		Prefer routine—Prefer variety	Ideas
			Values
		Uncurious—Curious	
		Conservative—Liberal	
A	Agreeableness, Friendly Compliance, Socialization vs. Antagonism	Ruthless—Soft hearted	Trust
		Suspicious—Trusting	Straightforwardness
		Stingy—Generous	Altruism
		Antagonistic—Acquiescent	Compliance
			Modesty
		Critical—Lenient	Tender mindedness
		Irritable—Good natured	
C	Conscientiousness, Will to Achieve, Constraint vs. Undirectedness	Negligent—Conscientious	Competence
			Order
		Lazy—Hardworking	Dutifulness
		Disorganized—Well-Organized	Achievement striving
			Self-discipline
		Late—Punctual	Deliberation
		Aimless—Ambitious	
		Quitting—Persevering	

Items are answered on a 5-point Likert scale ranging from strongly disagree to strongly agree.

Examples:
N: “I seldom feel nervous.”
E: “Sometimes I don’t stand up for my rights like I should.”
O: “I believe variety is the spice of life.”
A: “When making laws and social policies, we need to think aboutwho might be hurt.”
C: “I try to go to work or school even when I’m not feeling well”

Source: Costa Jr, P. T., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. Journal of personality disorders, 6(4), 343-359.

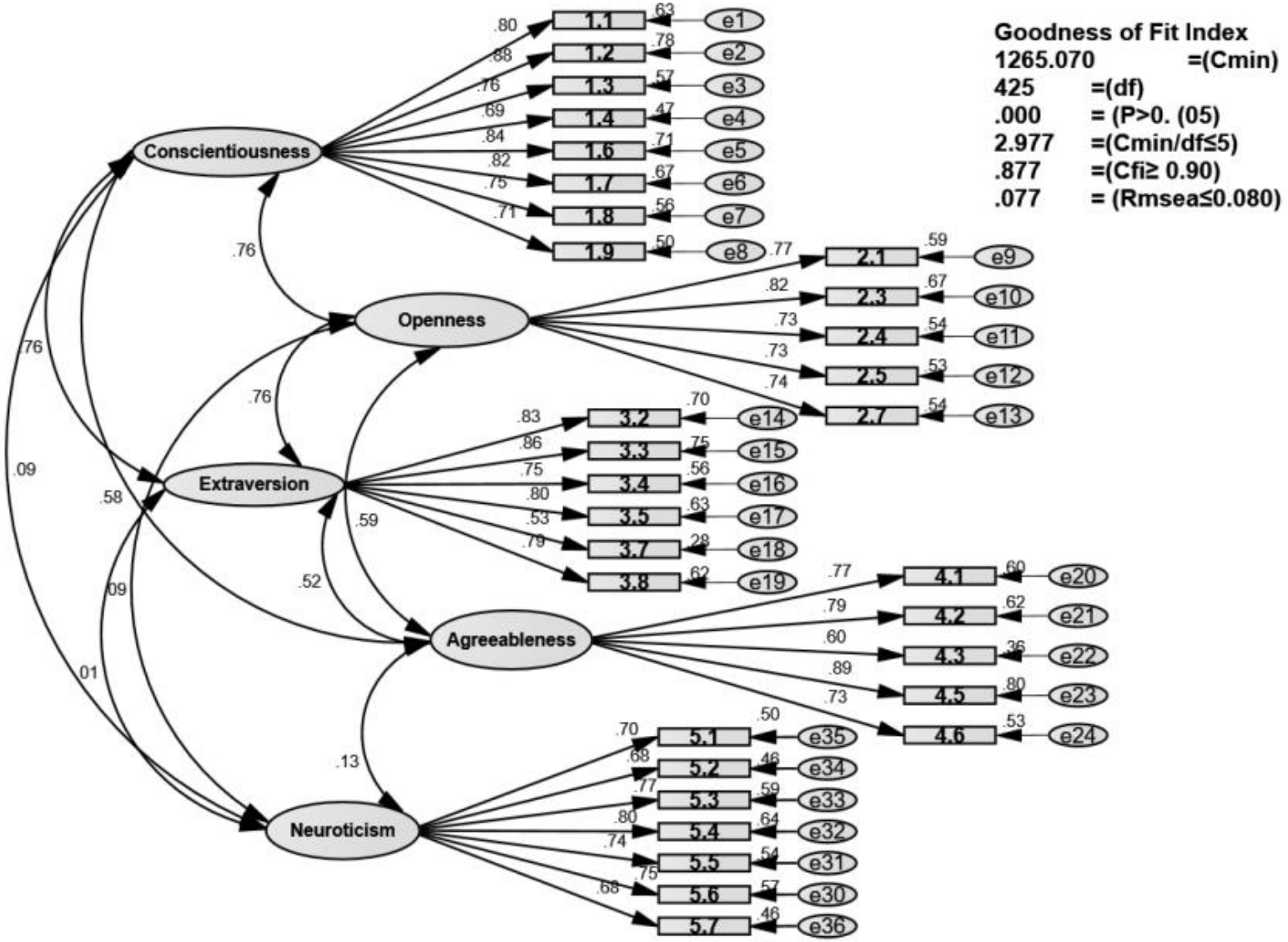
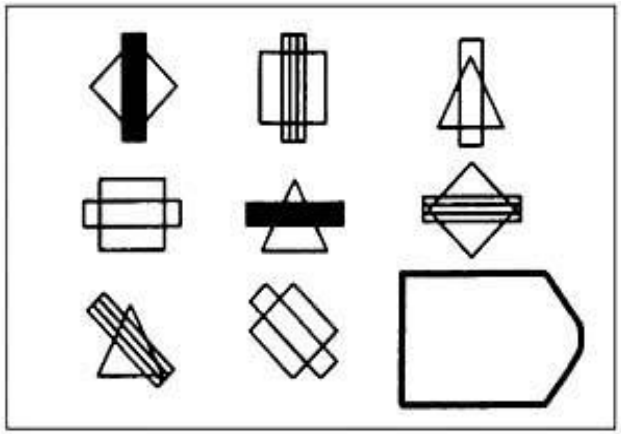


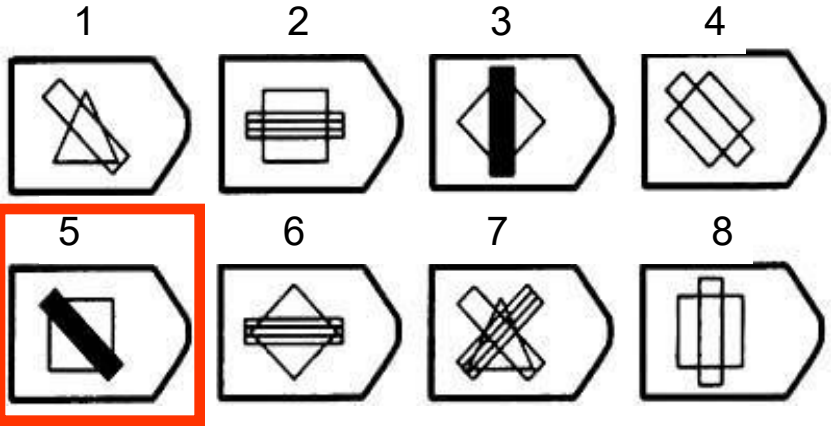
Figure from: Alharyout, A. O., & Abdullah, A. H. (2018). Confirmatory factor analysis of big five personality factors (CFA-BFPF). British Journal of Education, Learning and Development Psychology, 1(1), 29-36.

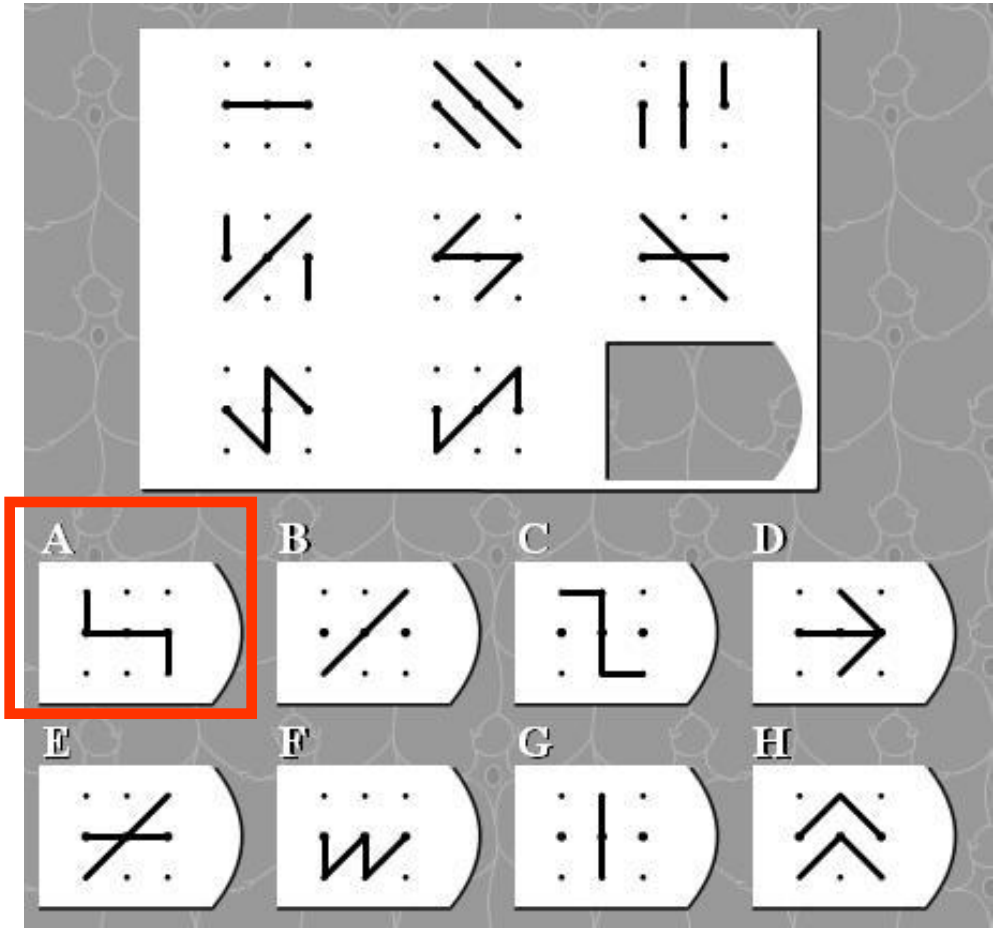
Intelligence

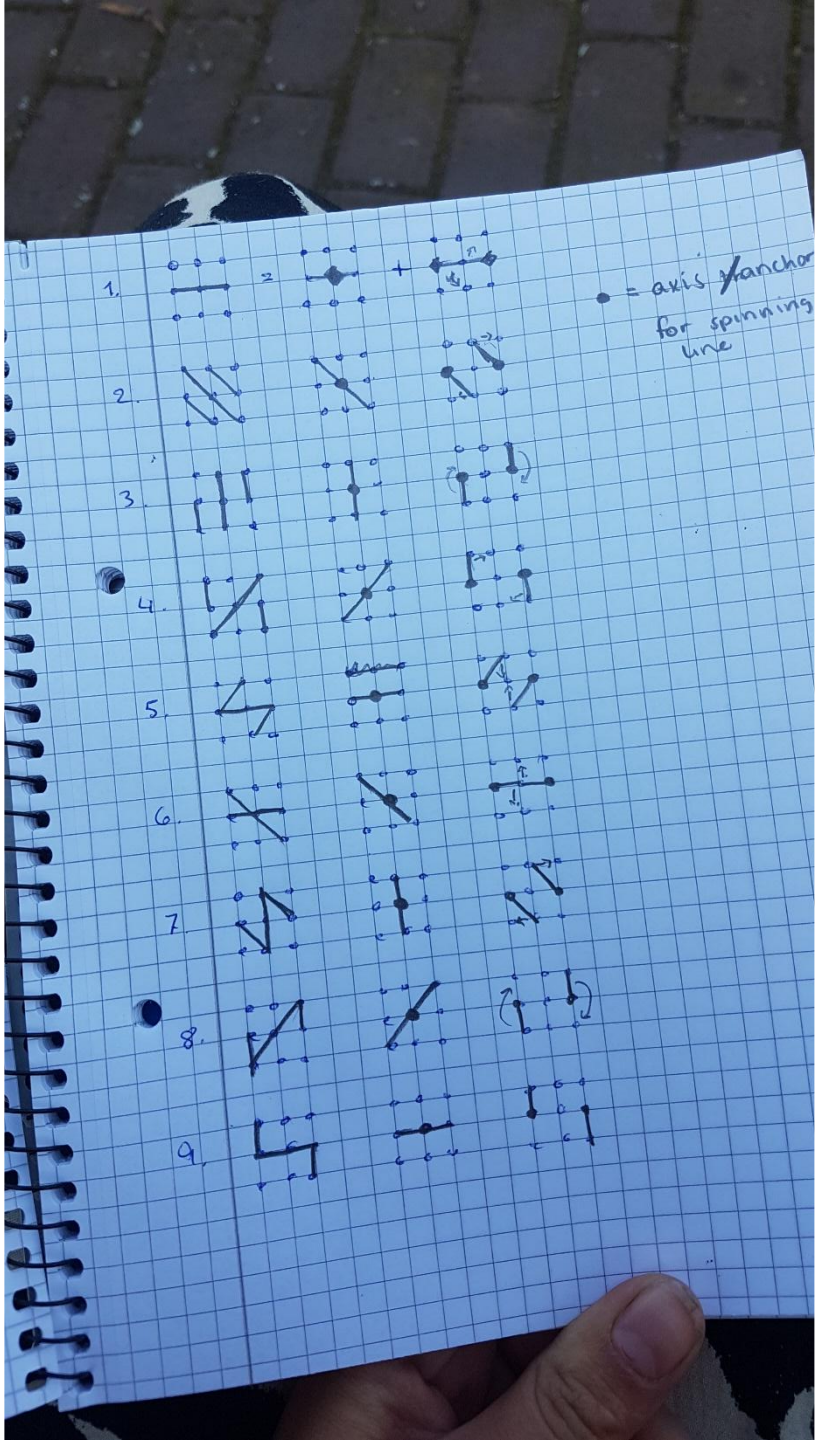
- Popular test for adults:
  "Wechsler Adult Intelligence Scale (WAIS)"
  - Four constructs:
    - Verbal comprehension
    - Perceptual organization
    - Working memory
    - Perceptual speed

Matrix reasoning (perceptual organization)

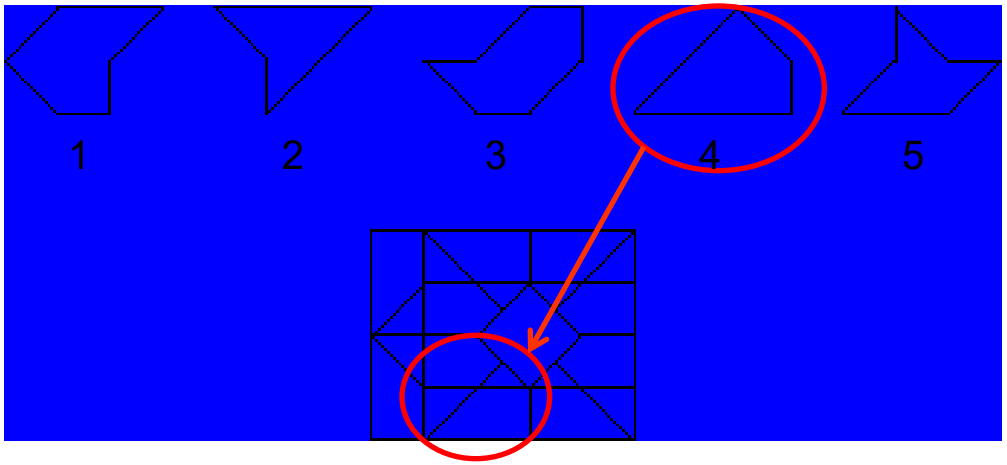




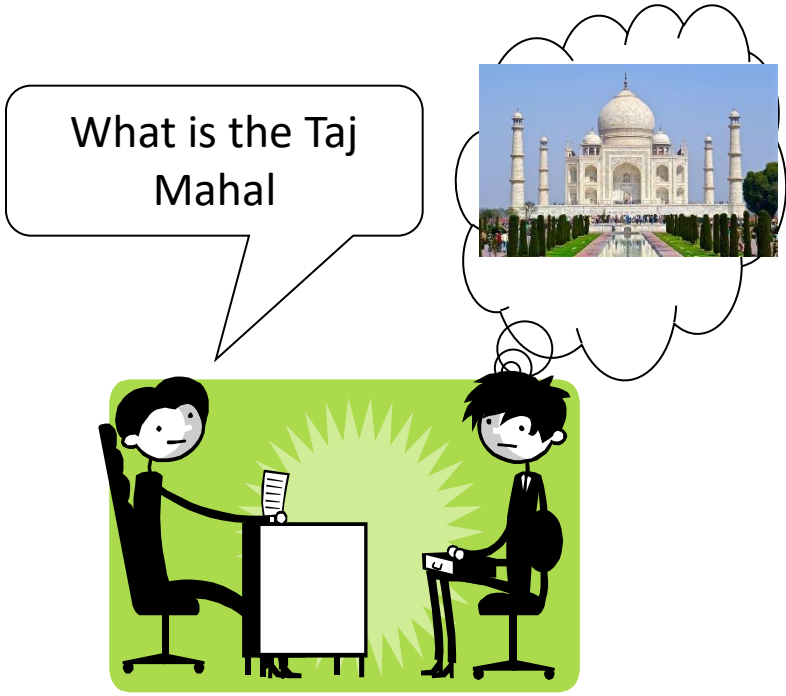



1.
2.
3.
4.
5.
6.
7.
8.
9.
= axis / anchor for spinning line

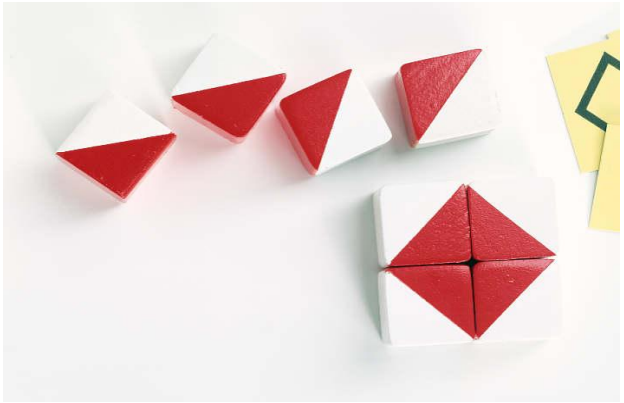
Hidden figures



General knowledge



Block design



Number sequences

2	4	6	8	…
0	1	3	7	…
1	4	27	256	….

Today

1. Measurement instruments in psychology
2. Problems
3. Quality of measurement in psychology

Objective vs Subjective

- Objective measures:
  - Observational measures
    - E.g., does someone come by bike
    - E.g., number of arithmetic items correct
    - E.g., does someone immediately make eye contact
  - Physical measures
    - Physiological measures (e.g.,: (f)MRI, heart rate)
    - Time, distance, etc
- Subjective measures:
  - Self-report
  - Expert judgement

Objective does not always mean “best”

- Alcohol breath test
  - “Hyperventilation for 20 seconds has been shown to lower the reading by approximately 32%”

- Polygraph (“lie detector”)
  - ‘Counter measures’ during the control questions
    - E.g., tighten your upper leg muscles

Objective does not always mean “best”

- From BNN National IQ test:

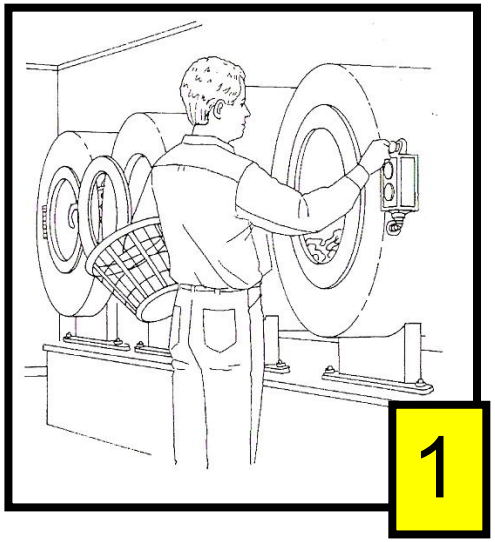
“Study the logo’s below for 60 seconds”

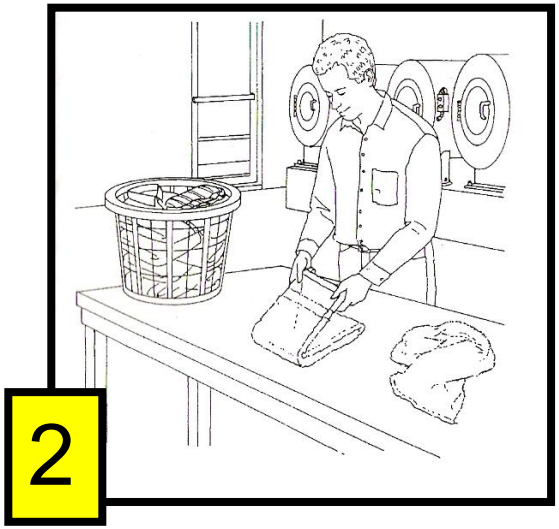
Objective does not always mean “best”

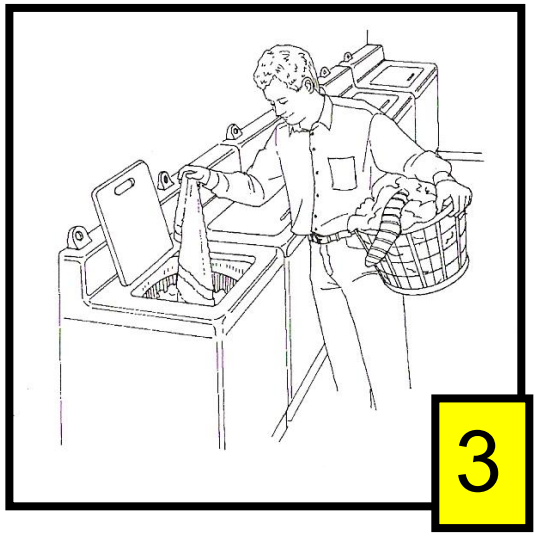
Which logo was left of the “proef konijnen” logo?

a) 3 op reis
b) De social club
c) Loverboys
d) Ranking the stars

Example from WAIS-III







Problems with Self-Report and Observers

- Key of the problem is subjectivity
- Various biases potentially influence the measurement outcome
  - Self-report:
    - All biases from lecture 1 (e.g., availability heuristic)
    - Social desireability, insecurity, poor self-knowledge, etc
  - Observers:
    - Also have these biases (e.g., confirmation bias)
- Question formulation→
- Answer options→

See chapter 6 of Morling for more examples of these

Question formulation

- Research by Loftus & Palmer (1974)
  - Participants looked at a small movie of a car accident
  - Question:
    - "What was the speed of the car when it ***hit*** the other car?"
    - "What was the speed of the car when it ***rammed*** the other car?"

Answer options

From van Heerden & Hoogstraten (1979)

1)	Yes	60 %
	Unsure	16 %
	No	24 %
2)	1	11 %
	2	27 %
	3	48 %
	4	14 %
3)	Very satisfied	24 %
	Satisfied	45 %
	Unsatisfied	16 %
	Very unsatisfied	15 %

In this investigation we want to determine how well you can guess the answers to questions that are not known to you. On each of the following pages you will find a Roman numeral and a group of possible answers. Sometimes these alternatives are words; in other cases they are figures or characters. The question, however, is missing. You are to circle one of the alternatives, the one that you think is correct. You can be sure that in principle each of the options is a fitting response. Please work through the items in the order given. Try not to skip pages, and circle one alternative per item only.[1]

4)	First	41 %
	Second	36 %
	Third	23 %
5)	True	60 %
	False	40 %
6)	A	28 %
	B	28 %
	C	27 %
	D	17 %
7)	Always	26 %
	Sometimes	46 %
	Never	28 %
8)	Agree	24 %
	Don’t care	60 %
	Disagree	16 %

Subjective and objective measures

- Both can be flawed, one is not automatically better than the other: it will depend on the attribute what kind of measure is most appropriate
- Most importantly: tests, questionnaires or other measurement instruments might not measure what we intend to measure with them.
- To evaluate *any* research, we need to know whether the measures it relies on are *valid*.

Today

1. Measurement instruments in psychology
2. Problems
3. Quality of measurement in psychology

Natural sciences

Example:

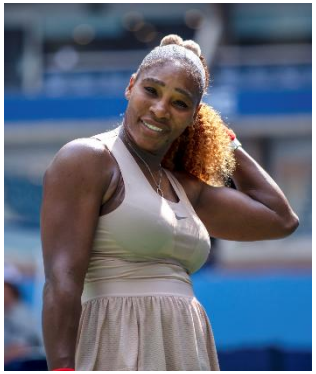
Heigth

Jordan Peele

Lady Gaga

Peter Starreveld

Dopey



Serena Williams

Nat ciences

Can we trust these measurements?

		height			
	Jordan	1.74	1.73	1.75	1.74
	Lady Gaga	1.70	1.71	1.69	1.69
	Peter	1.83	1.82	1.84	1.83
	Dopey	0.86	0.86	0.85	0.85
	Serena	1.74	1.73	1.74	1.73

Nat… Sciences

Can we trust these measurements?

		height	M1	M2	M3
	Jordan	1.74	1.73	1.75	1.74
	Lady Gaga	1.70	1.71	1.69	1.69
	Peter	1.83		1.84	1.83
	Dopey	0.86			
	Serena	1.74	1.73	1.74	1.73

Sure!
They are close to the real heights and they are consistent

Extraversion

At parties, within 10 minutes, I have everyones attention

Totally not applicable to me Totally applicable to me

1 2 3 4 5

I like to be the center of attention

Totally not applicable to me Totally applicable to me

1 2 3 4 5

I like to be in the company of others

Totally not applicable to me Totally applicable to me

1 2 3 4 5

Personality

Can we trust these measurements?

		Extraversion	M1	M2	M3
	Jordan	5	5	4	3
	Lady Gaga	4	4	3	5
	Peter	3	1	3	2
	Dopey	1	1	5	4
	Serena	4	5	5	1

Personality

Can we trust these measurements?

	Extraversion	M1	M2	M3
 Jordan	5	5	4	3
 Lady Gaga	4	4	3	5
 Peter	3	1	3	
 Dopey	1			
 Seren		5	5	1

Unsure!
-Do the scores correspond with conceptual variable?
-Can these scores be considered consistent?

Validity and Reliability

- Reliability
  - How consistent is the measure?
    - How much measurement error does the measure contain?
  - E.g.,: Do you get the same extraversion score every time you're tested?

- Validity:
  - Does the instrument measure what it is supposed to measure
  - E.g., Does a lie detector measure lying? Does an IQ test measure intelligence (or something else e.g., knowledge of american culture, or language)?

- Scores represent conceptual variable (validity)
- Scores are consistent (reliability)

nality

Ideal case (never happens)

	Extraversion	M1	M2	M3
 Jordan	5	5	5	5
 Lady Gaga	4	4	4	4
 Peter	3	3	3	3
 Dopey	1	1	1	1
 Serena	4	4	4	4

Personality

means

Reliable? Valid?

	Extraversion	M1	M2	M3	
 Jordan	5	5	4	6	**5**
 Lady Gaga	4	3	5	4	**4**
 Peter	3	2	2	4	**3**
 Dopey	1	1	2	0	**1**
 Serena	4	4	5	3	**4**

Personal

Reliable? Valid?

Reliable scores (reasonably consistent scores)

		Extraversion	M1	M2	M3	
	Jordan	5	5	4	6	5
	Lady Gaga	4	3	5	4	4
	Peter	3	2	2	4	3
	Dopey	1				1
	Serena	4	4			4

And valid (means correspond to conceptual variable)

Personality

Reliable? Valid?

Perfectly reliable measure (100% consistent scores)

	Extraversion	M1	M2	M3	
	Jordan	5	3	3	3
	Lady Gaga	4	5	5	5
	Peter	3	1	1	1
	Dopey	1	4	4	4
	Serena	4	2	2	2

Personality

Perfectly reliable measure (100% consistent scores)

		Extraversion	M1	M2	M3	
	Jordan	5	3	3	3	3
	Lady Gaga	4	5	5	5	5
	Peter	3	1	1	1	1
	Dopey	1			4	4
	Serena	4		2	2	2

But invalid!! (means don't correspond to conceptual variable)

Personality

Reliable? Valid?

Unreliable scores (not consistent scores)

		Extraversion	M1	M2	M3	
	Jordan	5	0	3	6	3
	Lady Gaga	4	6	3	6	5
	Peter	3	0	0	3	1
	Dopey	1	6	3	3	4
	Serena	4	1	0	5	2

Validity can't be judged

Reliability

- Are the scores consistent? -> reliable

Or do they vary a lot due to random noise (measurement error)? -> unreliable

- Reliability is generally a number between 0 and 1
  - E.g., 0.7 means that 70% of the variation you see between people is due to systematic differences and 30% is due to random measurement error.
  - 0 means that the scores are completely random

- Rules by COTAN (Dutch committee for evaluation of tests).
  - For high stakes use (detection of learning disabilities) .8/.9;
  - For group level use (comparing groups of students) .6/.7

  - Reliability of a lie detector: 0.40!

Reliability

How consistent are the scores (across repeated administrations, across raters, across items)?

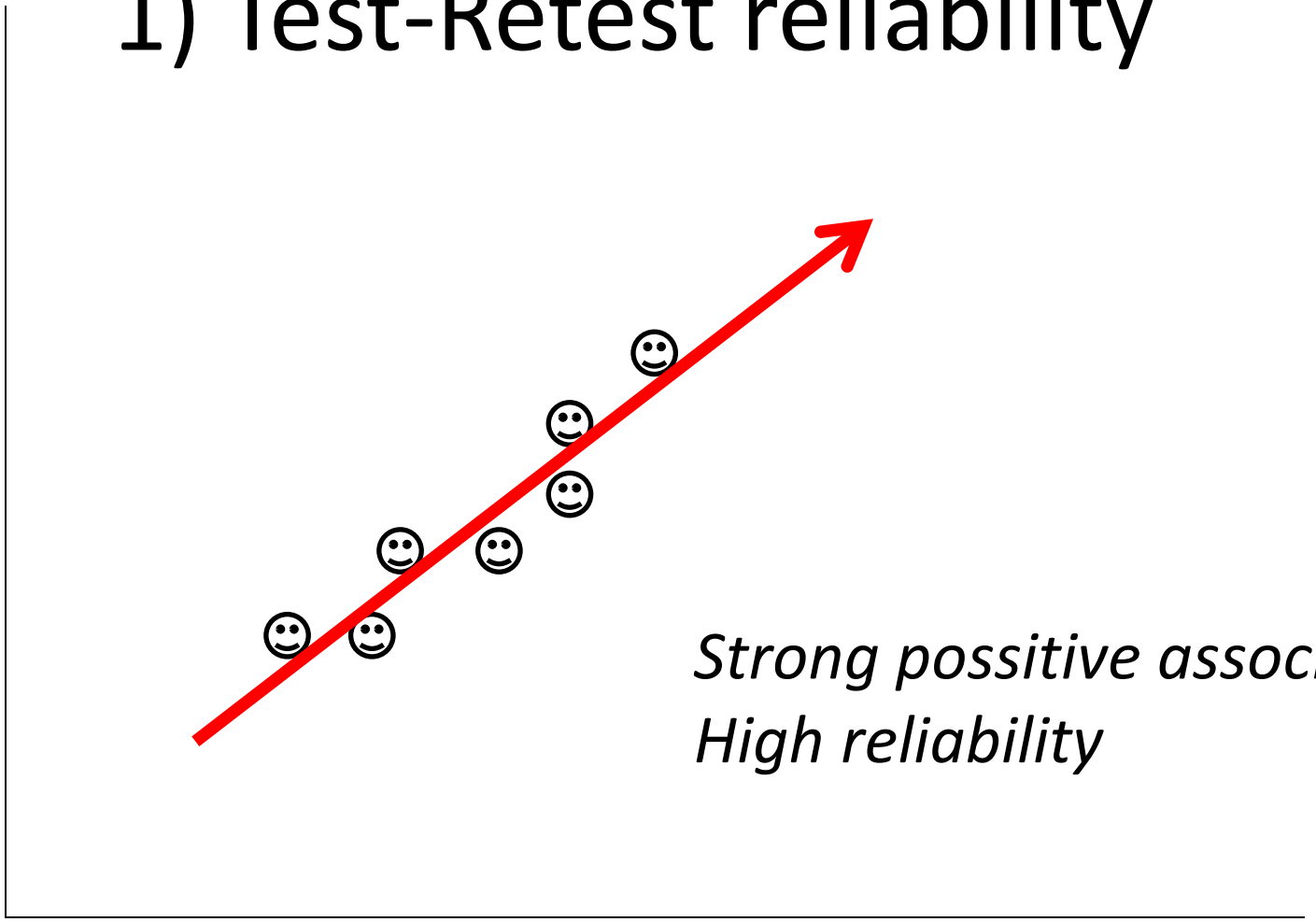
1. Test-retest reliability
2. Interrater reliability
3. Internal reliability

Reliability

How consistent are the scores (across repeated administrations, across raters, across items)?

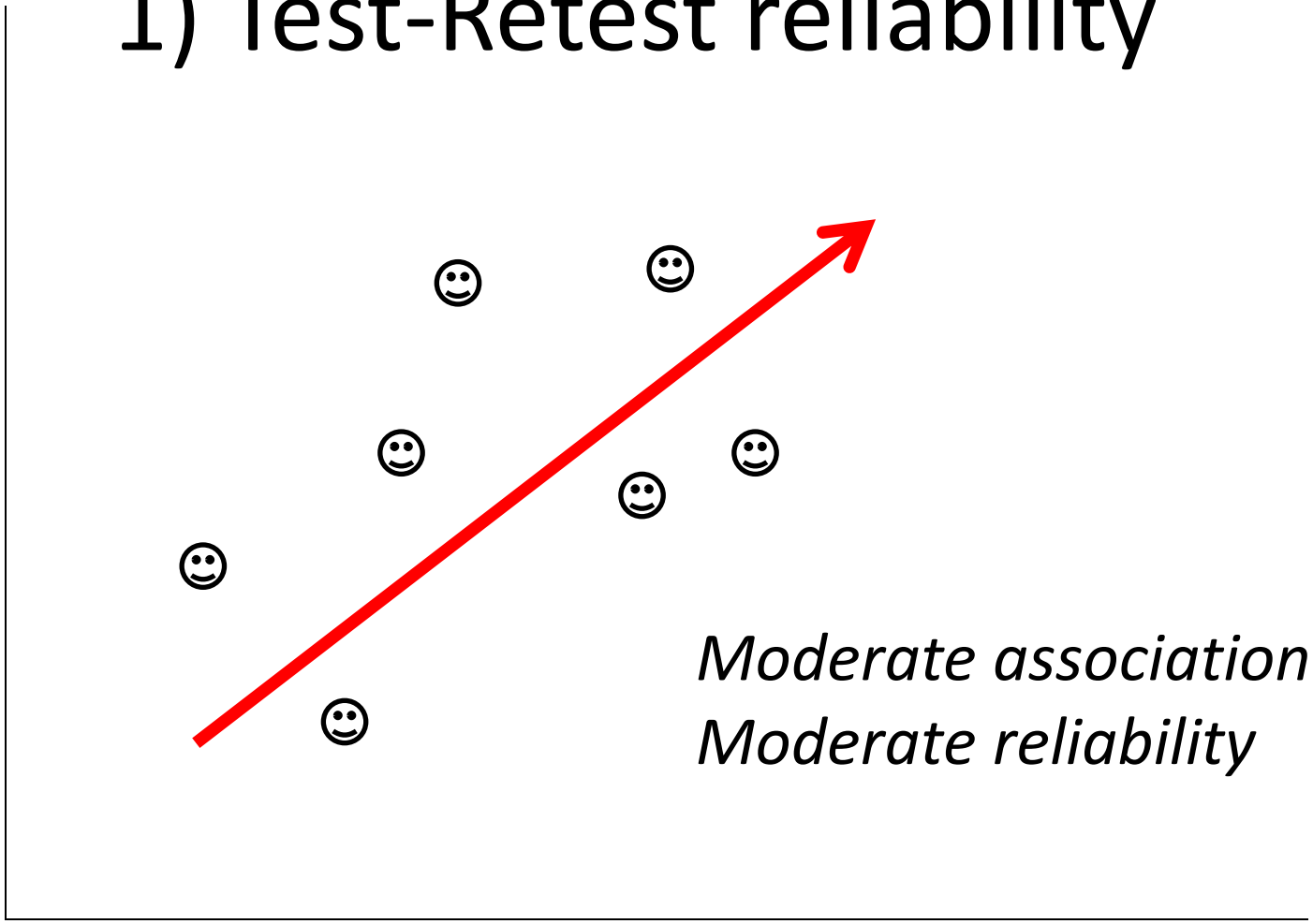
1. Test-retest reliability
   - Can be quantified by a correlation between test and retest
2. Interrater reliability
   - Can be quantified by a correlation between different raters
3. Internal reliability
   - Can be quantified by Cronbachs alpha which is based on the correlations between different items

1) Test-Retest reliability

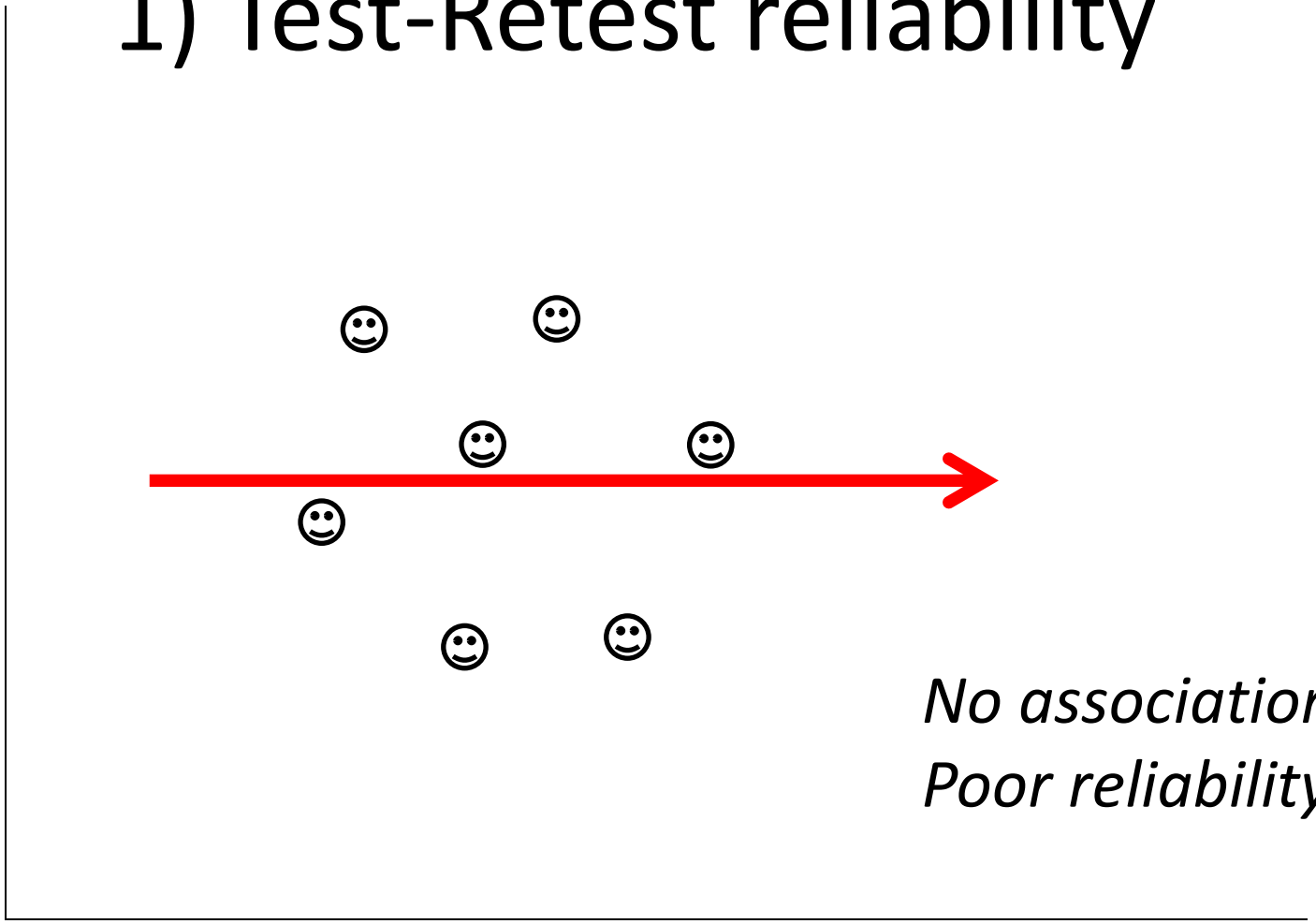


Strong possitive association:
High reliability

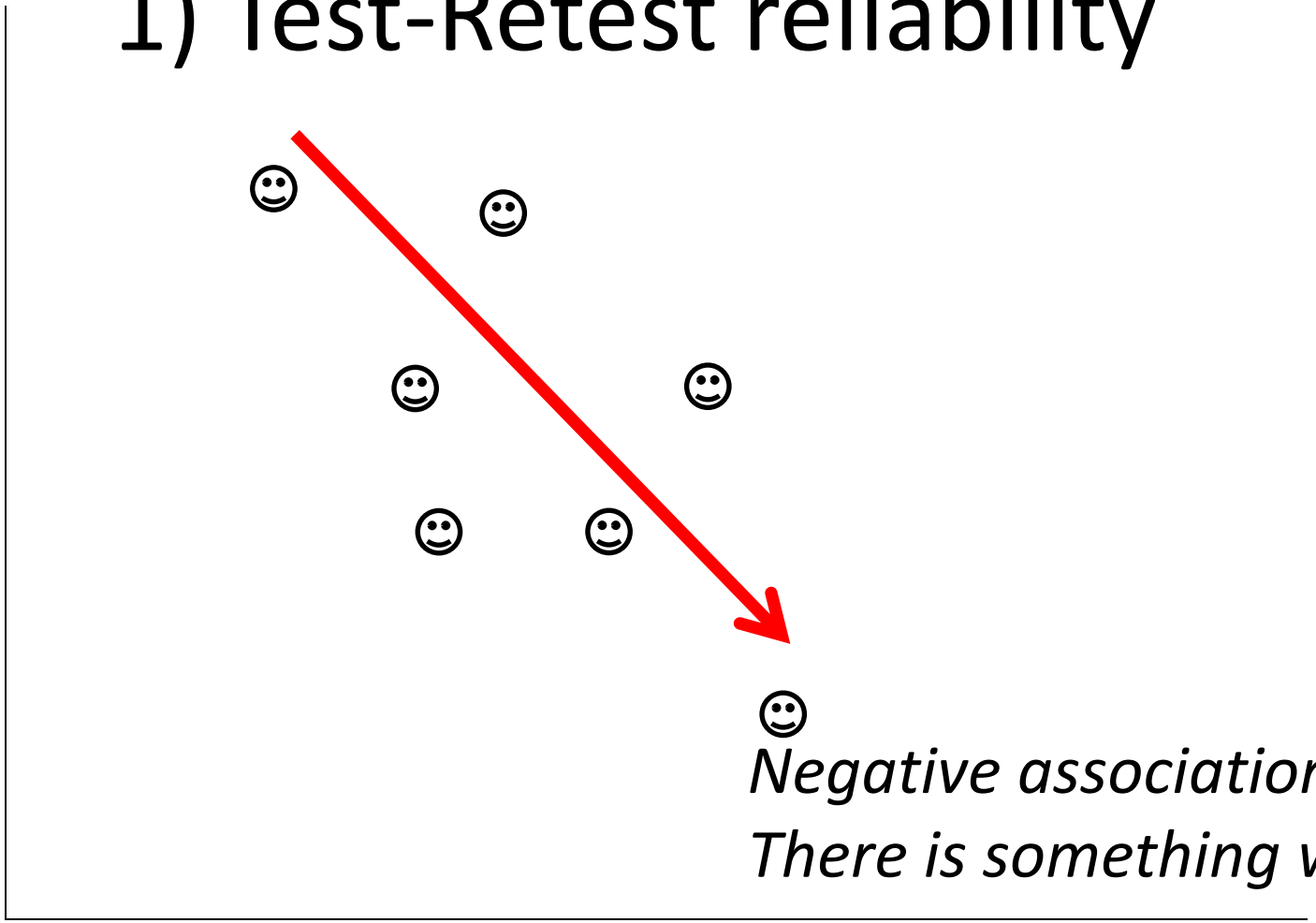
1) Test-Retest reliability



1) Test-Retest reliability



1) Test-Retest reliability



1) Test-Retest reliability

- Strength of the association between test and retest gives an indication of reliability

- Disadvantage: only works when the construct is *stable* between test and retest
  - No problem for variables that are assumed to be stable (e.g., intelligence and personality)
  - Problem for variables that we expect to fluctuate over time (e.g., mood, stress, depression severity)

Reliability

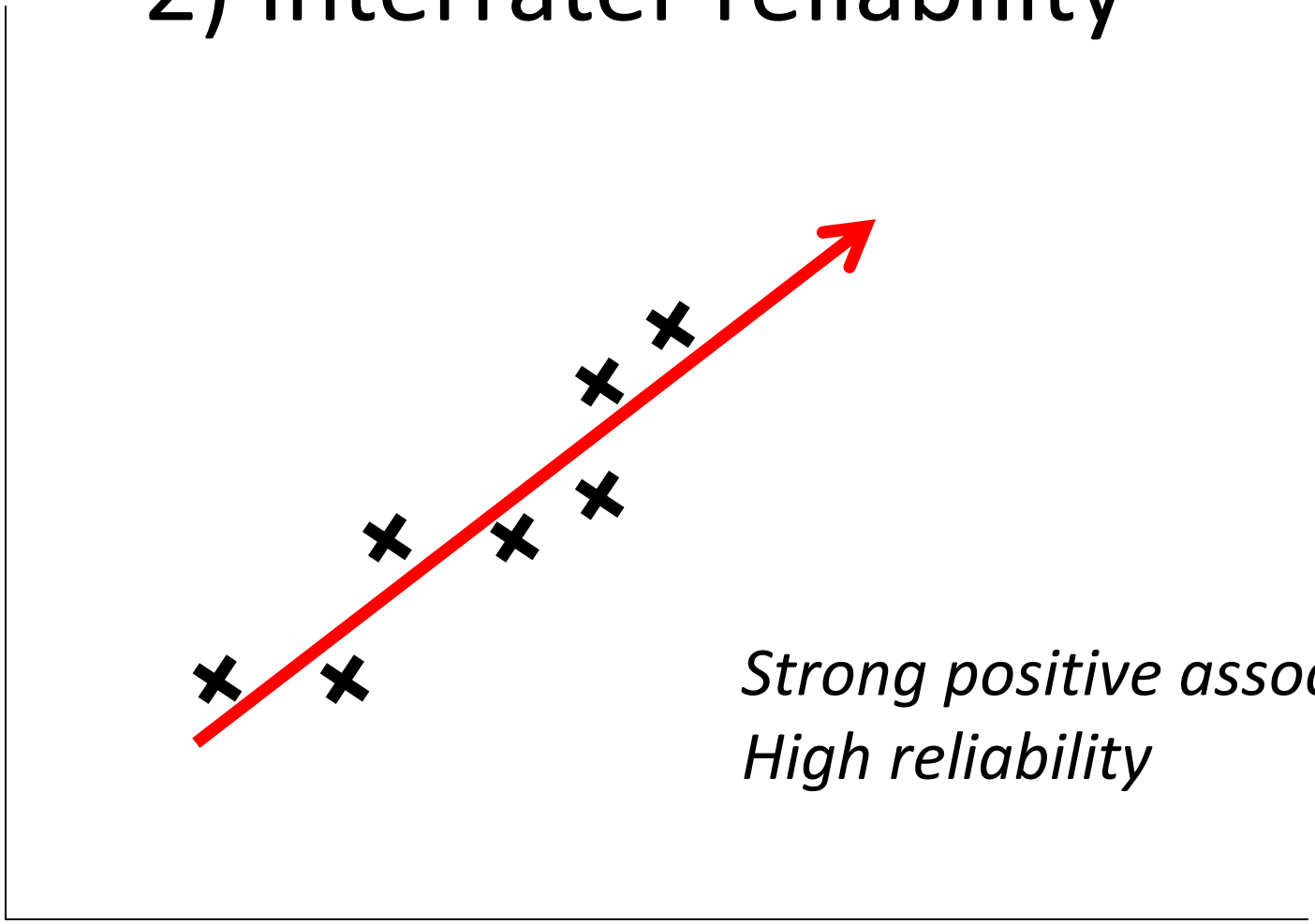
How consistent are the scores (across repeated administrations, across raters, across items)?

1. Test-retest reliability

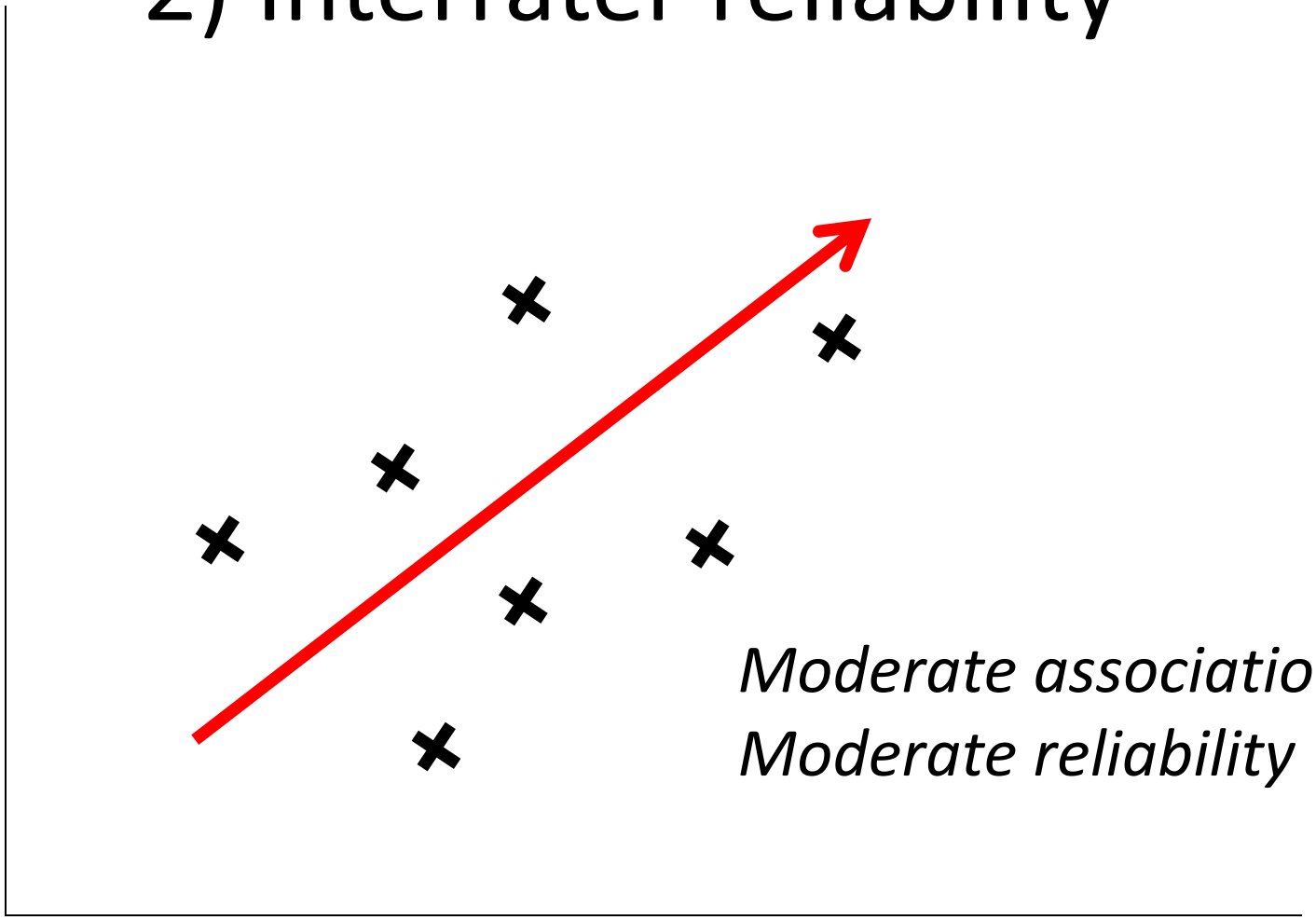
2. Interrater reliability

3. Internal reliability

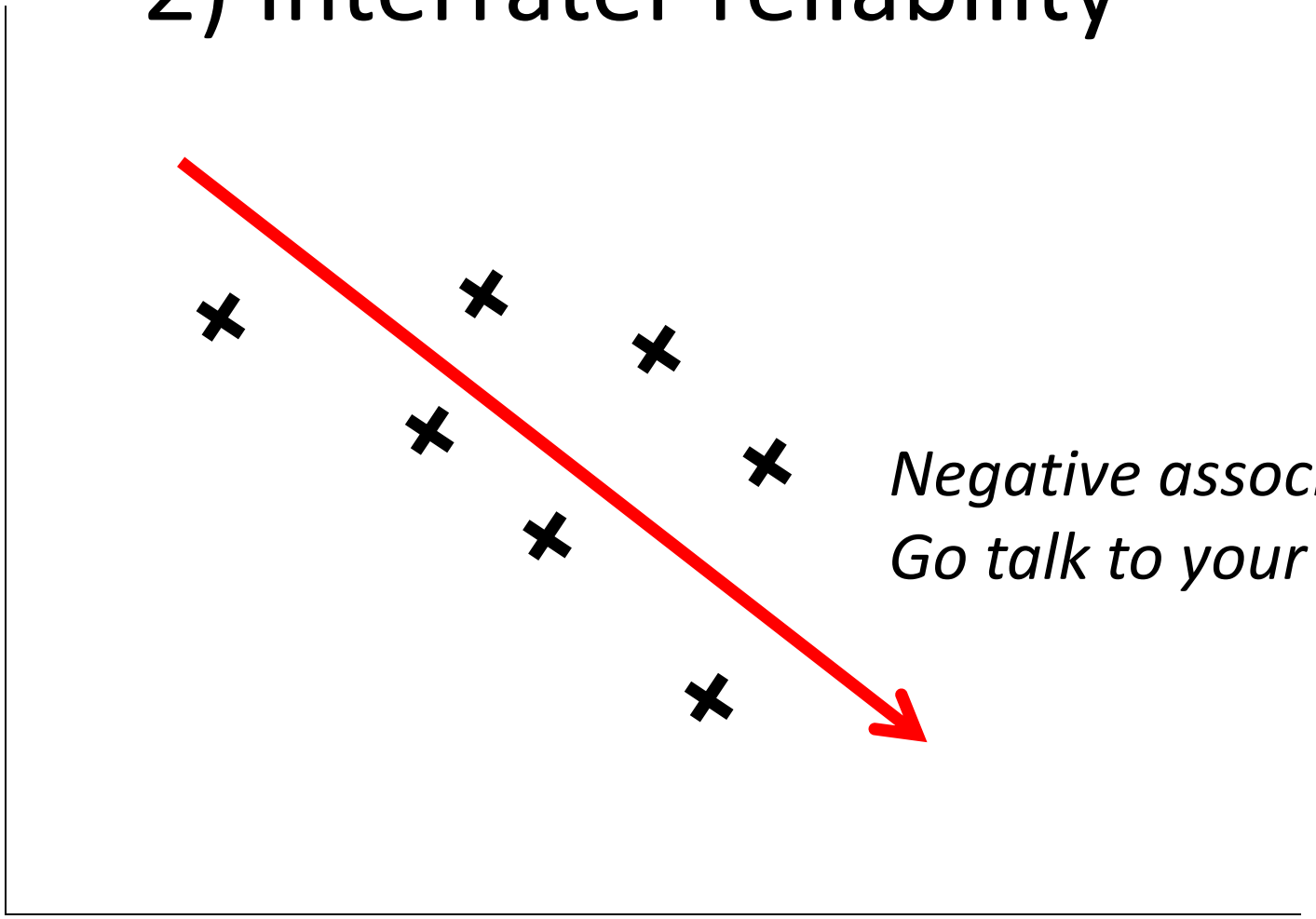
2) Interrater reliability



2) Interrater reliability



2) Interrater reliability



Negative association:
Go talk to your raters!

Reliability

How consistent are the scores (across repeated administrations, across raters, across items)?

1. Test-retest reliability

2. Interrater reliability

3. Internal reliability

3) Internal reliability

- Do the different items give the same impression about the person
  - E.g., someone who scores high (compared to others) on item 1, should also score high (compared to others) on item 2

- **Cronbach's alpha**: quantifies internal reliability
  - Based on the average association between the items
    - High average association: high reliability
    - Low average association: low reliability

  and number of items
    - More items: higher reliability
    - Less items: lower reliability

Checking your understanding

Suppose we have a test and we administer people's scores on two occasions and find a high correlation. We also find a high correlation between different raters, and Cronbach's alpha looks good as well!

Can we say the test is valid?

What if all these correlations are very low?

Validity

Does the instrument measure what it is supposed to measure

1) Face validity
2) Content validity
3) Criterion validity
4) Convergent validity
5) Discriminant validity

1) Face validity

→On the face of it, does the measure appear to be a plausible measure for the construct?

– E.g.,:
  - Low face validity: Shoe size for arithmatic ability
  - High face validity: Item 'At parties, within 10 minutes, I have everyones attention' for extraversion

– Face validity is subjective so do not consider face validity alone!

2) Content validity

- Does the measure cover all aspects of the construct of interest?
  - Common example: in the case of an IQ test, does it cover:
    - Working memory tasks
    - Spatial ability tasks
    - Verbal ability tasks
    - Etc
  - But also, a scale for measuring depression severity: does it consider only behavioral components or also affective components?
  - E.g., in an exam, are all main topics covered?

- Depends on your theory about the construct (so, also subjective)

3) Criterion validity

- Does the test correlate with an observable criterion (e.g., key behaviors)
- E.g.,
  - Does a happiness self report scale correlate with behavioral outcomes such as becoming ill and missing work
  - Does a sales performance aptitude test correlate with actual sales
  - Do depression severity scale scores differ between people diagnosed with depression and people who aren't

4) Convergent validity

- Is the test associated with other tests that measure the same (or a similar) construct
  - E.g., are the scores on the new extraversion measure related to existing extraversion measures?
    - Problematic if there are no well established tests

      E.g., “mindfulness”
  - E.g., are the scores on a depression scale correlated with well-being?

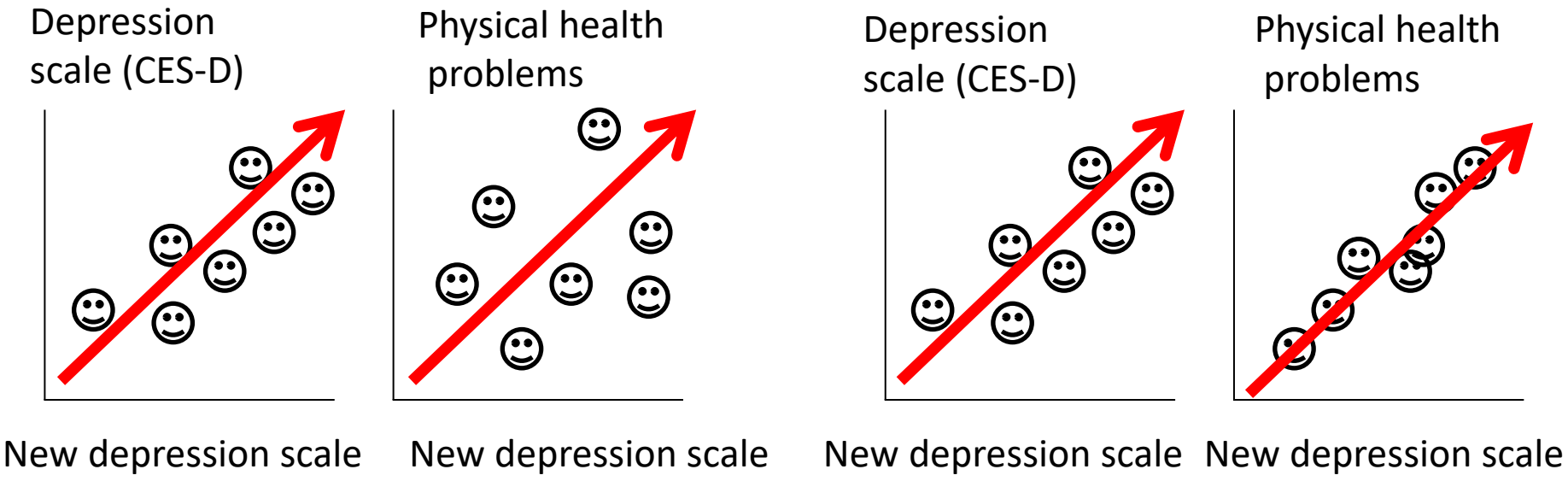
A single association is not convincing

- Tests scores can be associated to other variables but measure different things
  - E.g., ‘length’ is associated with ‘weight’, but a ruler and a scale measure different things

When a new depression scale correlates with existing depression scales or depression diagnoses, this is not sufficient: you also need discriminant validity

5) Discriminant validity

A test should not have a *stronger* association to a test that measures something else than with a test that is supposed to measure the same



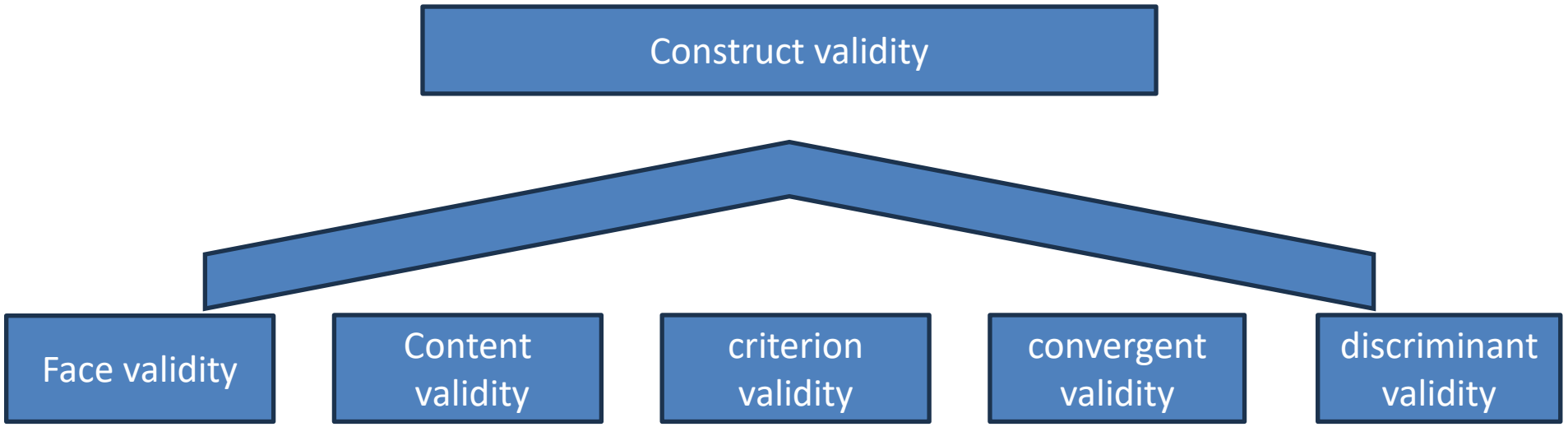
Discriminant validity

No discriminant validity

Construct validity

Does a test measure what it is supposed to measure?

- All validity evidence collected (discriminant validity, convergent validity, etc.) can contribute to this question



Validity

Does the instrument measure what it is supposed to measure

1) Face validity 2) Content validity	Subjective assessments
3) Criterion validity 4) Convergent validity 5) Discriminant validity	Empirical assessments: considering correlations

Only assessed when reliable!

Debate about Validity

DARTBOARD VALIDITY

Figure 1

The Dartboard Metaphor for Reliability and Validity

Submitted paper: *Dartboard Validity: A formal approach to quantifying item and test validity* by Mijke Rhemtulla, Anna Wysocki, and Riet van Bork

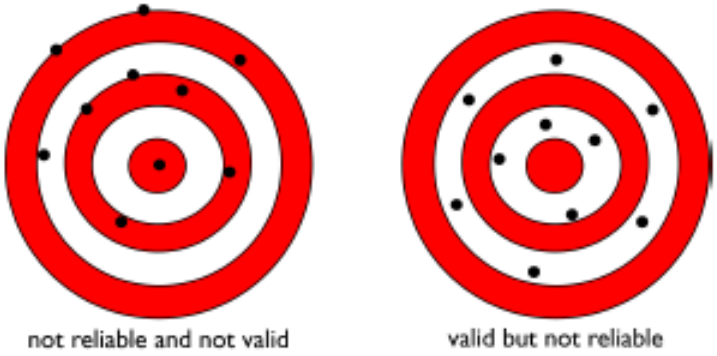
Reviewer: "Why do I give such a long winded example? Because although I really like this manuscript,
I strongly disagree with it. Should it be published in XX? Yes. Should a revision address my critique? I hope so."

Debate about Validity

DARTBOARD VALIDITY

Figure 1

The Dartboard Metaphor for Reliability and Validity



Submitted paper: *Dartboard Validity: A formal approach to quantifying item and test validity* by Mijke Rhemtulla, Anna Wysocki, and Riet van Bork

Reviewer: "Why do I give such a long winded example? Because although I really like this manuscript,
I strongly disagree with it. Should it be published in XX? Yes. Should a revision address my critique? I hope

Note: this disagrees with current definitions and the book, so for the exam: follow the book.
Why I show this example? Debate about what validity is and how to measure it is alive! You might contribute

Today

- Both objective and subjective measures have problems and so it is important to empirically study whether your measure indeed *measures what it is supposed to measure*. This is the question of construct validity.
- Reliability is the extent to which scores are consistent (across repeated administrations, across raters and across items)
- Both reliability and validity can be studied by looking at correlations
  - Reliability: correlations between test-retest, different raters or items.
  - Validity: correlations with a criterion or other measures that are expected to measure the same or expected to measure something else.

MC question

Tim created a questionaire with 8 items that all measure extraversion on a likert scale. However, it turns out that the item scores show only very weak correlations with each other. This is a problem for

a) internal reliability.
b) convergent validity.
c) discriminant validity.

MC question

Tim created a questionaire with 8 items that all measure extraversion on a likert scale. However, it turns out that the item scores show only very weak correlations with each other. This is a problem for

a) internal reliability.

b) convergent validity.

c) discriminant validity.

See Morling Ch.5, p.129-131.

And how well do you remember?

Which of the following is a problem of *experience* as a source of information?

a) There is no comparison group
b) People are swayed by a good story
c) People focus on the evidence they like best

And how well do you remember?

Which of the following is a problem of *experience* as a source of information?

a) There is no comparison group

b) People are swayed by a good story

c) People focus on the evidence they like best

This MC question is based on Morling Ch.2, p.26. The other answers represent problems with "intuition" as a source of information (p32-35).

The VSPA presents...

First Year's Weekend 2025!!!

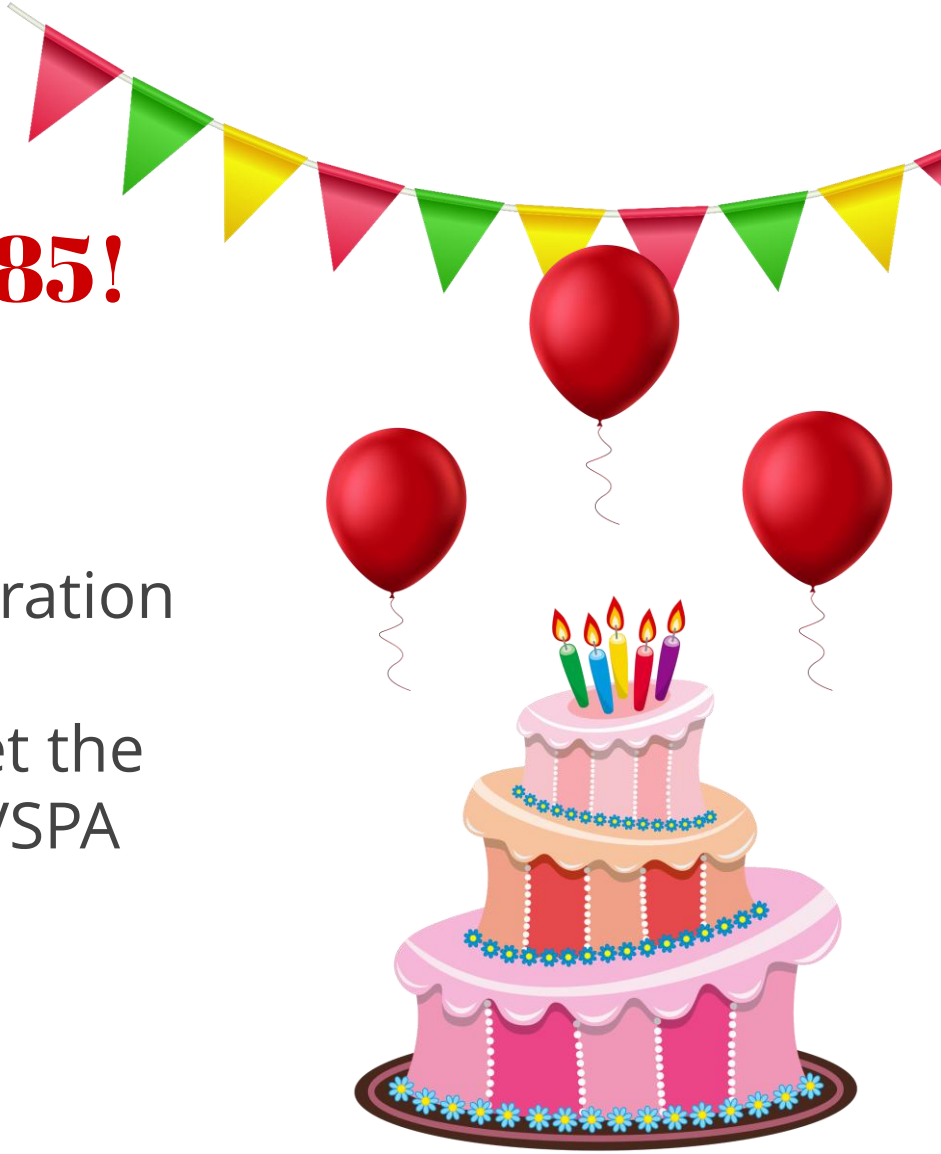
What to expect

- Bus transfer to and from the location
- Breakfasts, lunches, and dinners prepared by your mentors
- Plenty of games and activities throughout the whole weekend
- Two awesome parties!
- A chance to meet tons of your new peers
- Memories to last a lifetime

VSPA

The VSPA is turning 85!

- The theme of this FYW is… Birthday!
- Expect an extra large celebration
- More fun events to follow
- Now is your chance to meet the people you'll be seeing at VSPA events all year



Save the date

- September 26th-28th
- Secret location in southern NL
- Ticket sales coming soon!
- Follow the VSPA on Instagram and join the Whatsapp community to stay updated

@STUDYASSOCIATIONVSPA

VSPA

VSPA

WhatsApp community