Measurement

Research Methods for Human-Centered Computing

Measurement

Today's goal:

Teach how to measure subjective valuations (perceptions, experiences, intentions)

Outline:

- The theory of measuring things
- Using or adapting existing scales
- Developing new scales
- Pre-testing scales

Measuring things

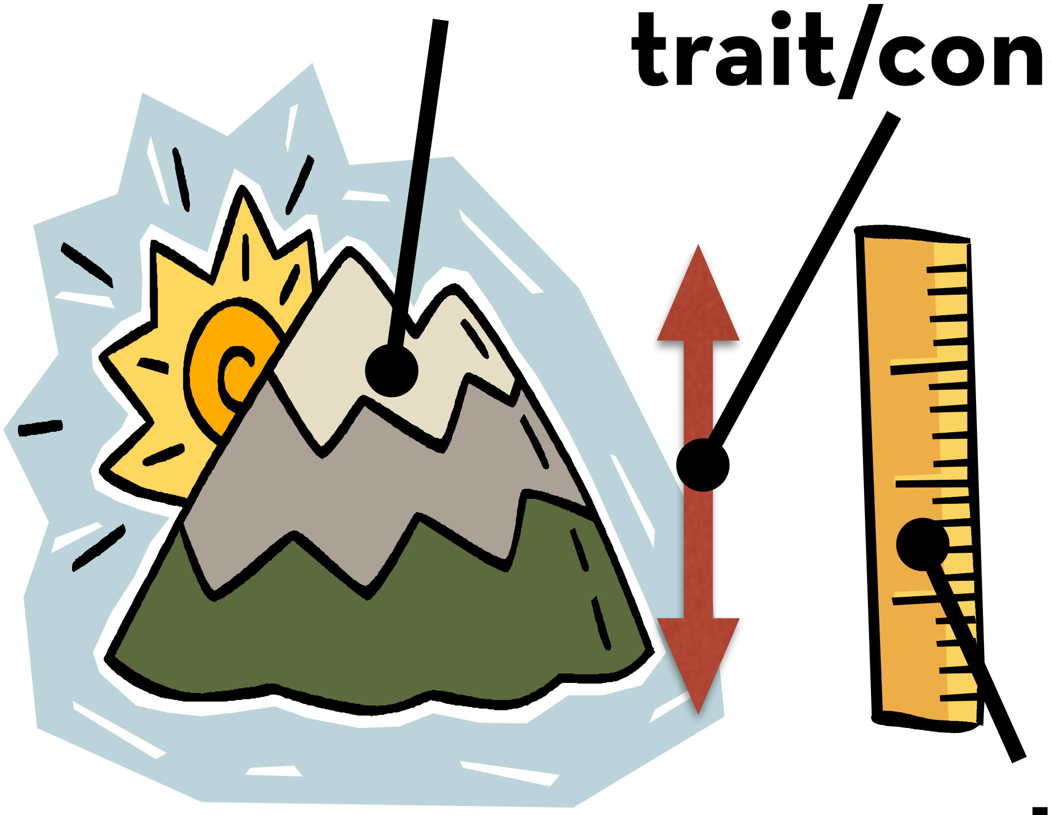
General theory

Measuring things

The quantification of a trait of an object

Using a method

On a scale



object

trait/concept

method + scale

Psychophysics

Some things cannot be observed directly, but their **experience** can be quantified by an observer

Examples:

- Temperature
- Loudness
- Pain

Psychometrics

The measurement of social and psychological concepts or traits

Rooted in the belief that these can be measured by asking questions (method)

Answers are an indirect observation on the concept/trait

Let’s try…

“To measure satisfaction, we asked users whether they liked the system (on a 5-point rating scale).”

Why is this bad?

Does the question mean the **same** to everyone?

- John likes the system because it is convenient
- Mary likes the system because it is easy to use
- Dave likes it because the outcomes are useful

A single question is not enough to establish **content validity**

We need a multi-item measurement scale

Scale: a collection of items, intended to reveal levels of a theoretical variable not readily observable by direct means

Why use a scale?

Objective traits can usually be measured with a single question

(e.g. age, income)

For subjective traits, single-item measurements lack **content validity**

Each participant may interpret the item differently

This reduces precision and conceptual clarity

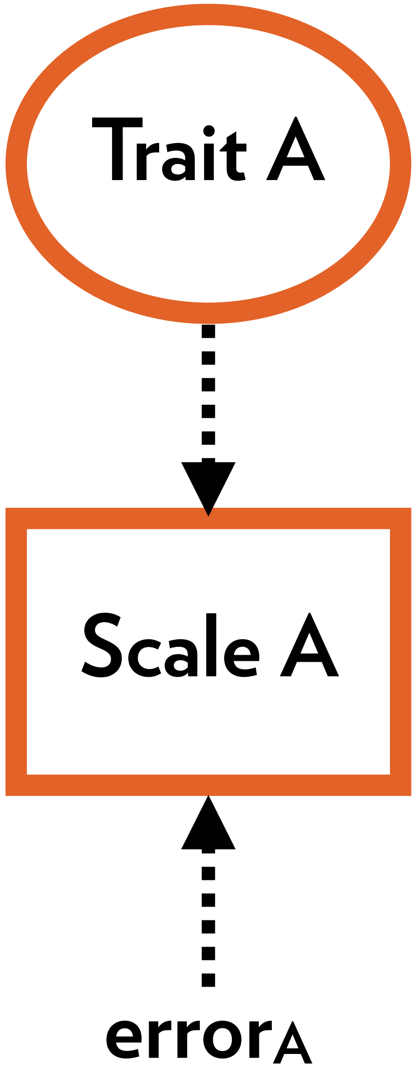
Accurate measurement requires a **shared conceptual understanding** between all participants and researcher

Latent variables

A scale is always an **imperfect** way of measuring a subjective trait

Our real goal is to measure the trait, not the scale

Scale = Trait + error



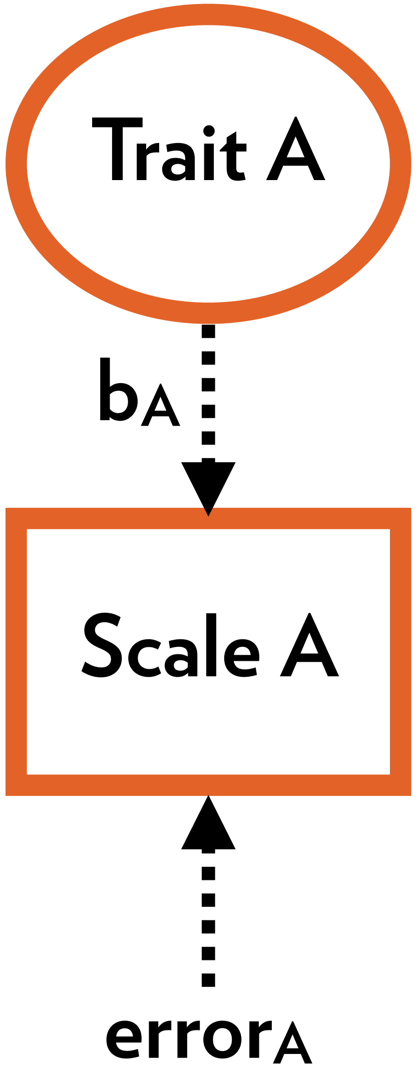
Latent variables

We can think of the traits as **latent** variables and the scales as **observed** variables

The trait **causes** my answers on the scale

Like a regression with an unobserved X

Scale A = a + b_A Trait A + error_A



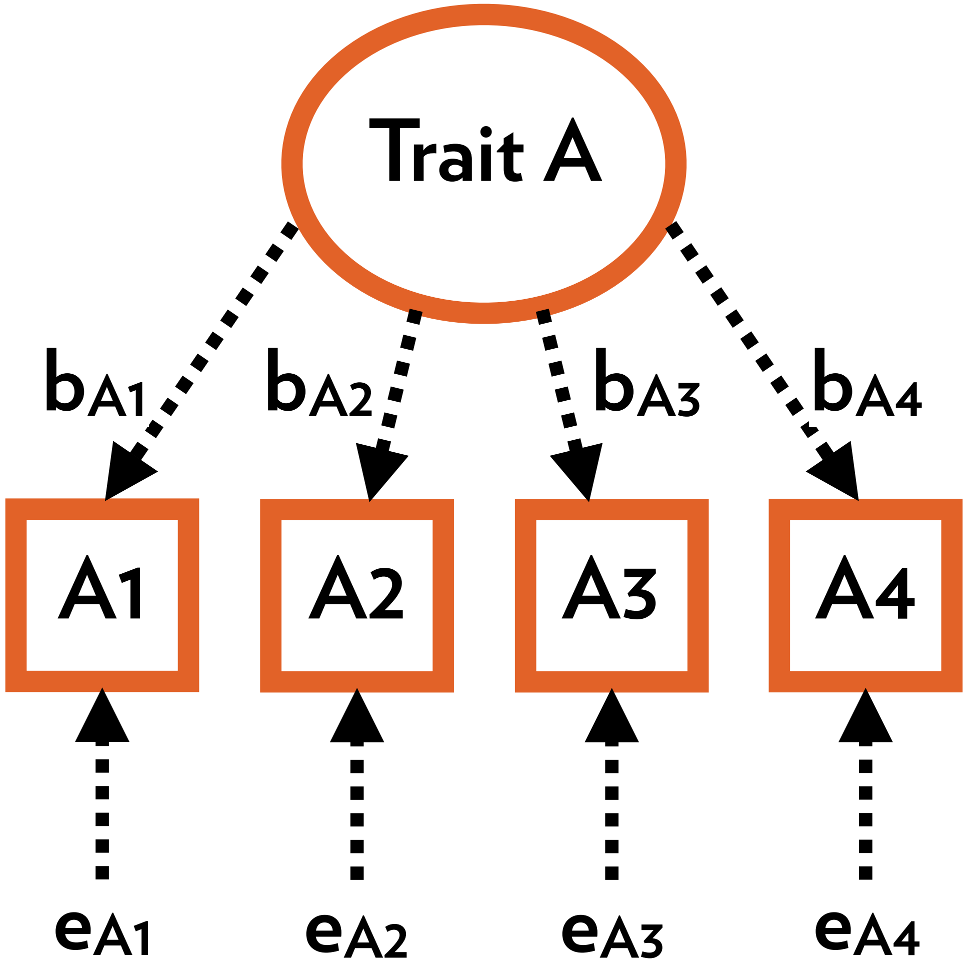
Latent variables

The R^2 of this regression determines how well we are measuring Trait A

  How do we get this R^2?

Trick: if you have multiple items, look at the **correlation** between the items

  Another reason to have multiple items!



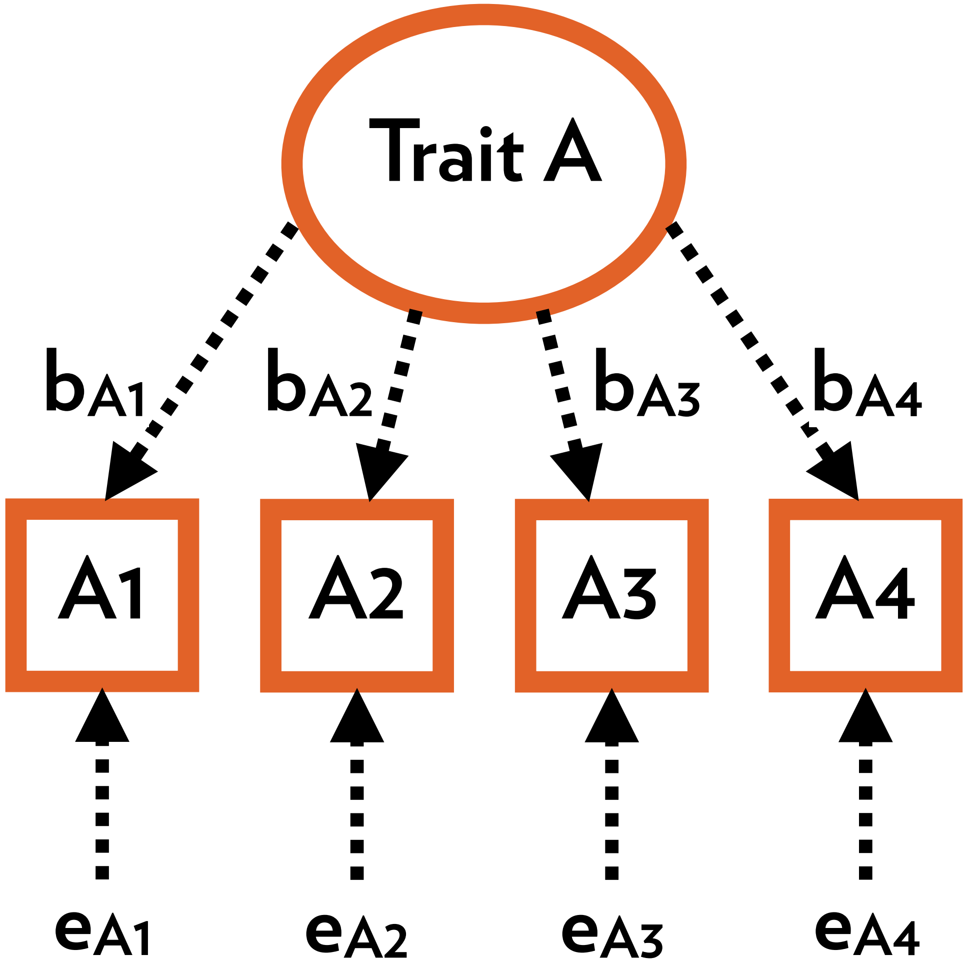
Latent variables

Let's say there are 4 items, each is correlated r = .64:

The b's are also called "loadings"

The e's are also called "uniqueness"

R^2 = 1-e is called "communality"

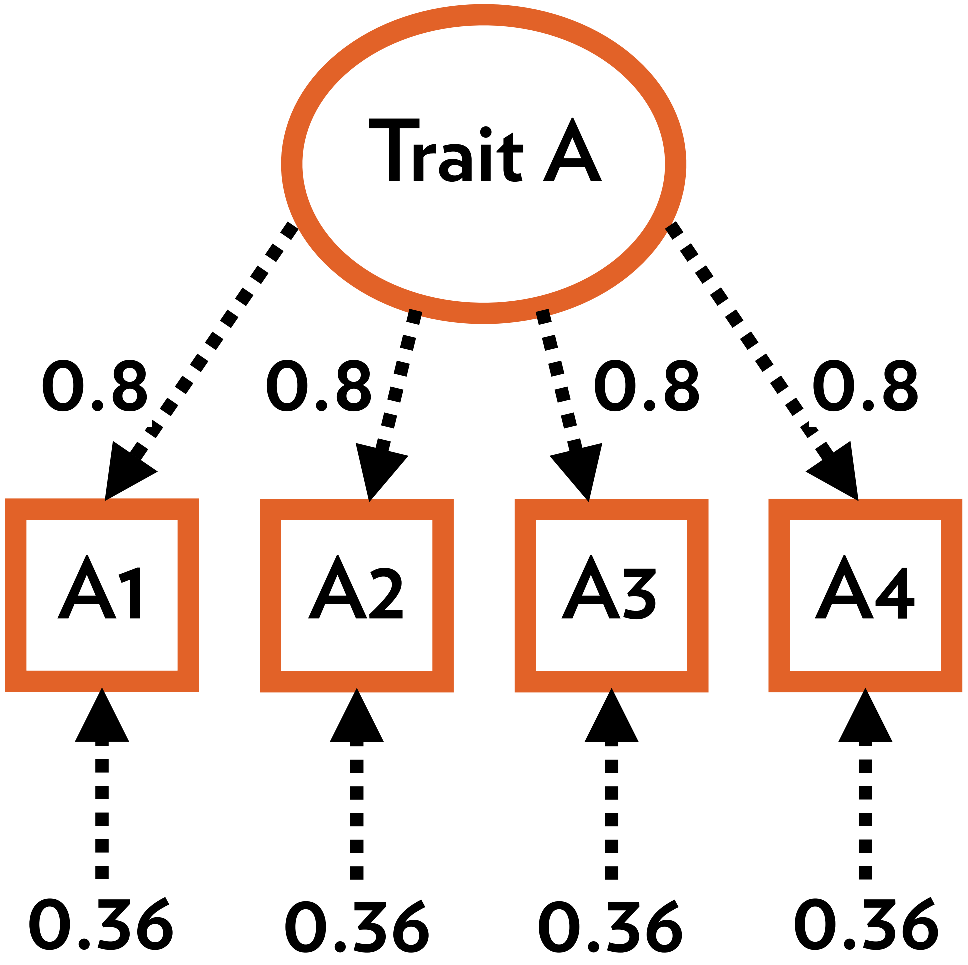


Latent variables

Fill in the numbers:

> To reconstruct the correlations, follow the paths!

(In M&E2 will do a version of this with multiple traits and unequal correlations)

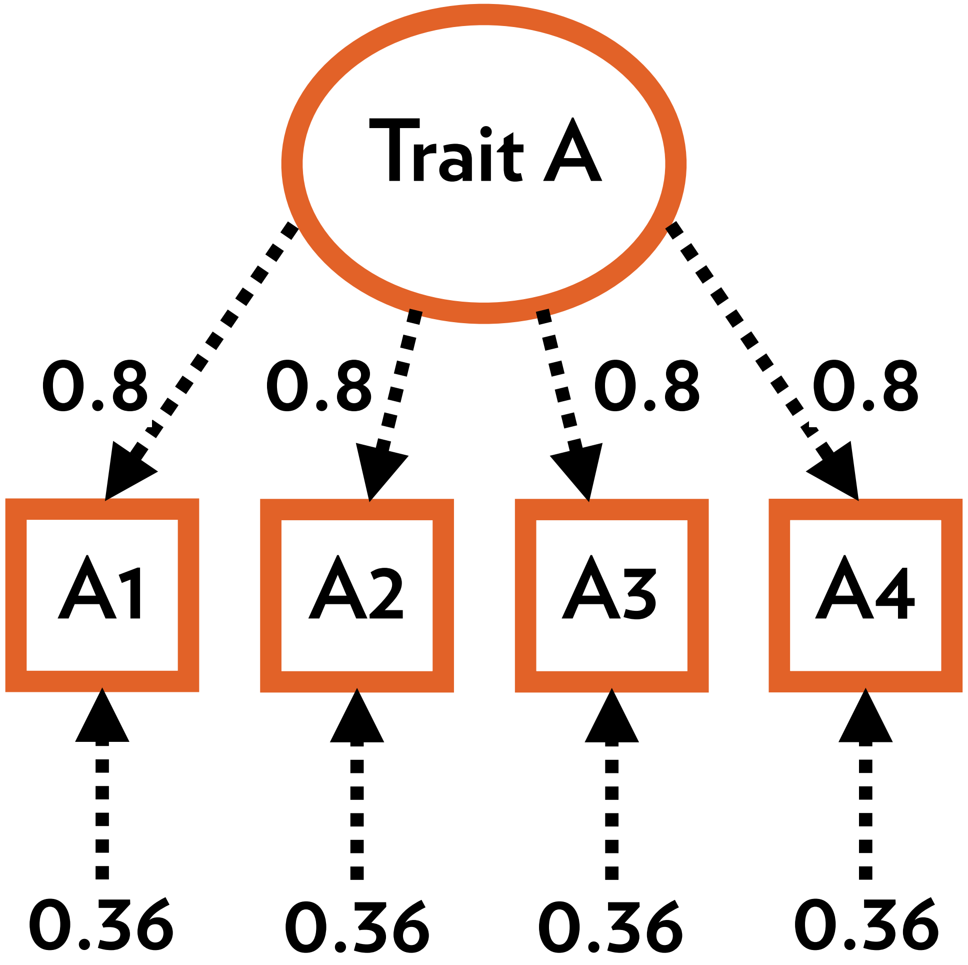


Reliability

Internal consistency is the extent to which the items measure the trait

Consistent scales have:

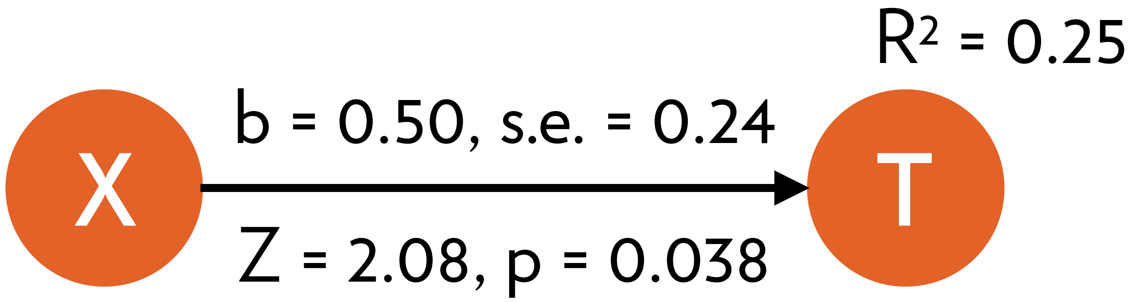
- Low uniquenesses
- High communalities
- High loadings
- High correlation between items



Problems...

Any regression coefficient will be **attenuated** by the reliability of the scale!

Take for instance this X, which potentially explains 25% of the variance of trait T...



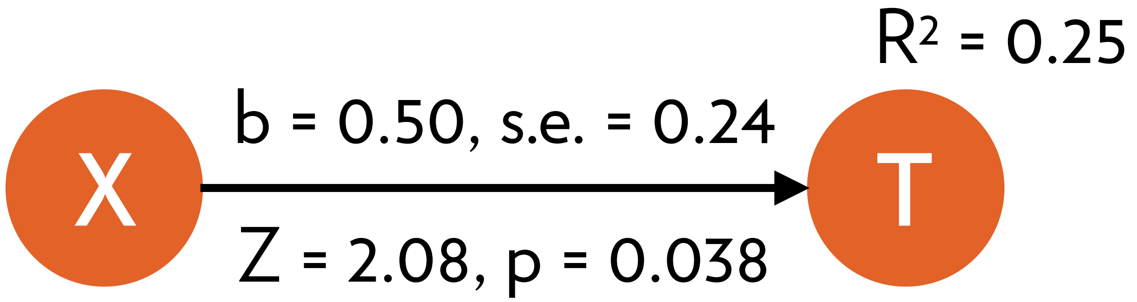
Problems…

However, trait T is measured by 4-item Scale S, which has loadings of 0.8 instead of 1.0

X only explains 16% of the variance of S!

…and the effect is non-significant!

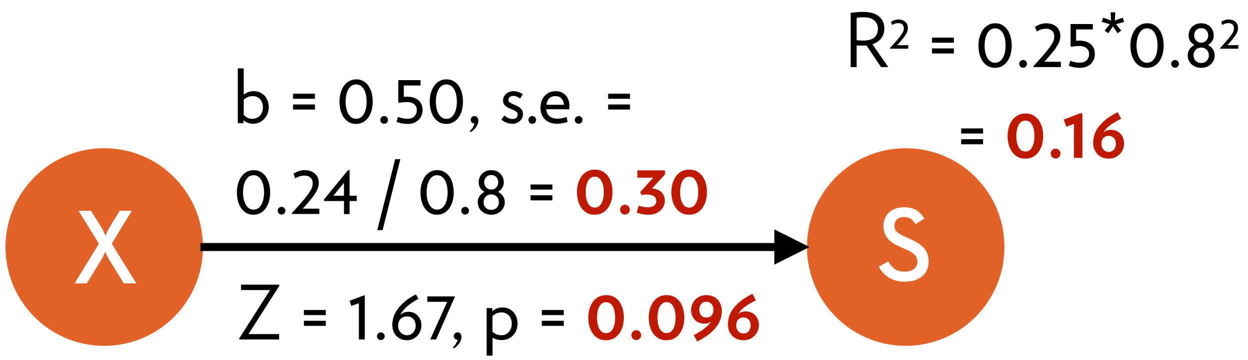
Higher reliability = more statistical power



$R^2 = 0.25$

b = 0.50, s.e. = 0.24

Z = 2.08, p = 0.038



$R^2 = 0.25*0.8^2$

= **0.16**

b = 0.50, s.e. = 0.24 / 0.8 = **0.30**

Z = 1.67, p = **0.096**

Solution!

In M&E2, we will learn **Structural Equation Modeling**, a method that retains the power regardless of the reliability of the measurement scales!

Scale development

Existing, adapted, and new scales

Use existing scales

Why?

- Constructing your own scale is a lot of work
- "Famous" scales have undergone extensive validity tests
- Ascertains that two related papers measure exactly the same thing

Finding existing scales:

- In related work (especially if they tested them)
- The Inter-Nomological Network (INN) at inn.theorizeit.org

Create new scales

When?

- Existing scales do not hold up
- Nobody has measured what you want to measure before
- Scale relates to the specific context of measurement

How:

- Adapt existing scales to your purpose
- Develop a brand new scale (see next slides!)

Adapting scales

Information collection concerns:	System-specific concerns:
It usually bothers me when websites ask me for personal information.	It bothered me that [system] asked me for my personal information.
When websites ask me for personal information, I sometimes think twice before providing it.	I had to think twice before providing my personal information to [system].
It bothers me to give personal information to so many websites.	n/a
I am concerned that websites are collecting too much personal information about me.	I am concerned that [system] is collecting too much personal information about me.

New scales: steps

1. Create a concept definition
2. Generate items
3. Determine the response format
4. Pre-Test the items
5. Include validation items
6. Administer the scale to a development sample
7. Evaluate the items
8. Optimize scale length

Concept definition

Start by writing a good concept definition!

A concept definition is a careful explanation of what you want to measure

Examples: leadership

"Leadership is power, influence, and control" (objectivish)

"Leadership is status, respect, and authority" (subjectivish)

"Leadership is woolliness, foldability, and grayness" (nonsensical, but valid!)

Concept definition

A good concept definition…

- …is grounded in and bounded by substantive theories
- …has an adequate level of specificity
- …makes it unambiguously clear what the concept is supposed to mean
- …is the foundation for a shared conceptual understanding

Note: This is an equality relation, not a causal relation

- Power, influence, control == leadership
- Not: power, influence, control —> leadership

Specificity

The specificity depends on your goal, e.g., compare:

“The world is run by the few people in power, and there is not much the little guy can do about it.”

“I feel like what happens in my life is determined by powerful others.”

“Having regular contact with my physician is the best way for me to avoid illness.”

“If I see my doctor regularly, I am less likely to have problems with <my condition>.”

Concept definition

If a concept becomes “too broad”, split it up!

- e.g. you could create separate concept definitions for power, influence, and control

If two concepts are too similar, try to differentiate them, but otherwise integrate them!

- e.g. “attitude towards the system” and “satisfaction with the system” are often very similar
- avoid situations where items fit with both scales

Creating items

E.g. Concept: "Leadership = status, respect, authority"

- Find a way to measure these aspects in a leader

Each item should reflect the concept, not just part of it

Write first, be critical later

- End up with 10-15 reasonable items, after removing the obviously bad ones

Redundancy is good! This supports the detection of the common concept

Creating items

Items should be somewhat different, but not just semantically

Bad:

“In my opinion, pet lovers are kind.”

“In my estimation, pet lovers are kind.”

Good:

“I think that people who like pets are good people.”

Creating items

The respondent does not have to be the measured object!

- E.g. for leadership, one could ask employees to rate their supervisor

Example items:

- "My supervisor is an admirable person."
- "I am more important than my supervisor."

Creating items

For objective concepts, you need to ask objective questions

E.g. behavior: "I do X" rather than "I like X"

Otherwise an exam could ask a single question:

> Do you believe that your understanding of the course materials is sufficient to pass this course?
>
> ( ) yes   ( ) no

Good items…

Use both positively and negatively phrased items

- They make the questionnaire less “leading”
- They help filtering out bad participants
- They explore the “flip-side” of the scale

The word “not” is easily overlooked

Bad: “The results were not very novel.”

Good: “The results felt outdated.”

Downside: negatively phrased items often perform poorly

Good items...

- Avoid asking respondents to say “yes” in order to mean “no”
  - Bad: Do you favor or oppose not allowing the state to raise taxes without a 60% approval rate?
  - Good: Do you favor or oppose requiring a 60% approval rate in order to raise taxes?
- Shoot for a low reading level
  - Bad: “Do you find the illumination of your work environment sufficient to work in?”
  - Test reading level: www.read-able.com

Good items…

Soften the impact of objectionable questions

Bad: “I do not care about the environment.”

Good: “There are more important things than caring about the environment.”

However, don’t make your questions too “mild”

Good items...

Avoid double-barreled questions

Bad: “The recommendations were relevant and fun.”

This could be two items, or even two scales!

Use appropriate time referents

E.g. cybercrime awareness: before or after the crime occurred?

Solution: explicitly mention the time referent in your question

Good items...

Avoid vague qualifiers or fuzzy words with an ambiguous meaning

Bad: “On the weekends I get down with my friends.”

Good: “I take the car for short distances (less than 7 miles).”

Avoid check-all-that-apply questions

Bad: “Which of the following cybercrimes have you been a victim of?” (check all that apply)

Good: “Have you been a victim of ___________?” (yes - no)

Response format

Most common types of items: binary, 5- or 7-point scale

- Binary items are less precise, but easier to answer
- Having more that 7 categories is rarely useful
- Exception: using a visual analog scale for very subtle effects

Usually, we want to measure the **extent** of the concept

- Examples on the next slides…

Response format

Likert scale:

Question preamble: (To what extent) do you agree or disagree with the following statement?

Question: <the statement>

Answer categories:

- completely disagree, disagree, somewhat disagree, neutral, somewhat agree, agree, completely agree
- no - yes

Response format

Question: (How often) do you … ?

- never, rarely, occasionally, frequently, very frequently
- no - yes

Question: How important is … to you? / Is … important to you?

- unimportant, mostly unimportant, somewhat important, rather important, very important
- no - yes

Response format

Question: How would you rate… ?

- very poor, poor, somewhat poor, neutral, somewhat good, good, very good

Question: How likely are you to … / I would likely …

- very unlikely, unlikely, somewhat unlikely, neutral, somewhat likely, likely, very likely
- false - true

Response format

Sometimes, the answer categories represent the item

Based on what I have seen, FormFiller makes it ________ to fill out online forms.

- easy - - neutral - - difficult
- simple - - neutral - - complicated
- convenient - - neutral - - inconvenient
- effortless - - neutral - - daunting
- straightforward - - neutral - - burdensome

Response format

Decide on whether you want a “neutral” option

- or: “neither agree nor disagree”
- Most often, this results in better scales

“Undecided” and “neutral” are not the same thing

- Bad: disagree - somewhat disagree - undecided - somewhat agree - agree
- Good: disagree - somewhat disagree - neutral (or: neither agree nor disagree) - somewhat agree - agree

Response format

Examples:

http://www.gifted.uconn.edu/siegle/research/instrument%20reliability%20and%20validity/Likert.html

Examples

Satisfaction:

- In most ways FormFiller is close to ideal.
- I would not change anything about FormFiller.
- I got the important things I wanted from FormFiller.
- FormFiller provides the precise functionality I need.
- FormFiller meets my exact needs.

(completely disagree - disagree - somewhat disagree - neutral - somewhat agree - agree - completely agree)

Examples

Satisfaction (alternative):

- Check-it-Out is useful.
- Using Check-it-Out makes me happy.
- Using Check-it-Out is annoying.
- Overall, I am satisfied with Check-it-Out.
- I would recommend Check-it-Out to others.

(completely disagree - disagree - somewhat disagree - neutral - somewhat agree - agree - completely agree)

Examples

Satisfaction (another alternative):

I am ________ with FormFiller.

- very dissatisfied - - neutral - - very satisfied
- very displeased - - neutral - - very pleased
- very frustrated - - neutral - - very contended

Attention checks

Always begin with clear directions

- Ask comprehension questions about the directions

Make sure your participants are paying attention!

- “To make sure you are paying attention, please answer somewhat agree to this question.”
- “To make sure you are paying attention, please do not answer agree to this question.”
- Repeat certain questions
- Test for non-reversals of reverse-coded questions

Scale development

Existing, adapted, and new scales

Testing items

Expert discussion

Ask experts to:

- provide feedback on the concept definition
- rate how relevant each item is
- evaluate the clarity and conciseness of each item
- suggest additional items

Testing items

Card sorting (both experts and users)

Steps:

- Print your scales, cut out the questions
- Ask the expert/user to sort the questions into groups
- Ask them to explain what they think each of the resulting groups is supposed to measure (concept definition)
- Remove/revise items that are in the wrong group, revise scales that got an incorrect definition

Testing items

Think-aloud testing

Ask users to

- read each question aloud (note any readability issues)
- give an answer to the question (note any doubts)
- explain the question in their own words (note any comprehension problems)
- explain their answer (note whether their answer reflects the intended construct)

(points 2 and 4 are not always possible)

Validation items

Optionally, test for social desirability

- Participants who score high on this scale are more likely to try to please you

Measure additional scales to establish concurrent validity

- E.g., measure compassion if you think that it should correlate with your altruism scale

First test of a scale

Administer your scales to a development sample

- Target N: 5 times the number of items

It is okay if the sample is not the target population, and sometimes scales can be tested without a system (or in the context of a different system)

- As long as the participants are expected to have some value on the latent trait to be measured

Evaluate the items

Using Factor Analysis, see M&E2!

Optimize length

Final scale should have least 3 (but preferably 5 or more) items per scale

Developing items involves multiple iterations of testing and revising

- First develop 10–15 items
- Then reduce it to 7-10 through discussions with domain experts and comprehension pre-tests with test subjects
- You may remove 1-2 more items after the first test
- You may remove another 1-2 in the final analysis

Assignment

Measurement scale assignment

Assignment

Goal: develop a subjective measurement scale for your group project.

Note:

- this is an individual assignment
- you are encouraged but not required to use the scale in your project
- you will develop the scale but not validate it

Instructions

Write down a subjective concept

something that can be measured using multi-item measurement scales

Examples (feel free to use, or develop your own):

- perceived company trustworthiness
- perceived agent human-likeness
- perceived conversation quality
- perceived mentor support
- perceived stressfulness of a task

Instructions

Write a clear a concise concept definition

- 1-2 paragraphs
- should make it 100% clear what you are trying to measure

Develop 10-15 items

- take the best practices for item development into account
- keep an eye on the reading level of your questions
- select an appropriate response format (7-point agreement is most common)

Instructions

You can reuse/adapt items from existing scales

- make sure to provide a reference!

Please do not reuse an existing scale verbatim

- that's normally allowed, but not in this assignment

Submit the assignment as a PDF on Canvas by Oct 16