Psychometrics

Measuring subjective valuations

Intro

Today's goal:

Teach the general idea of measuring subjective valuations (perceptions, experiences, intentions)

Outline:

- The theory of measuring things
- Latent variables
- Reliability
- Validity

Measuring things

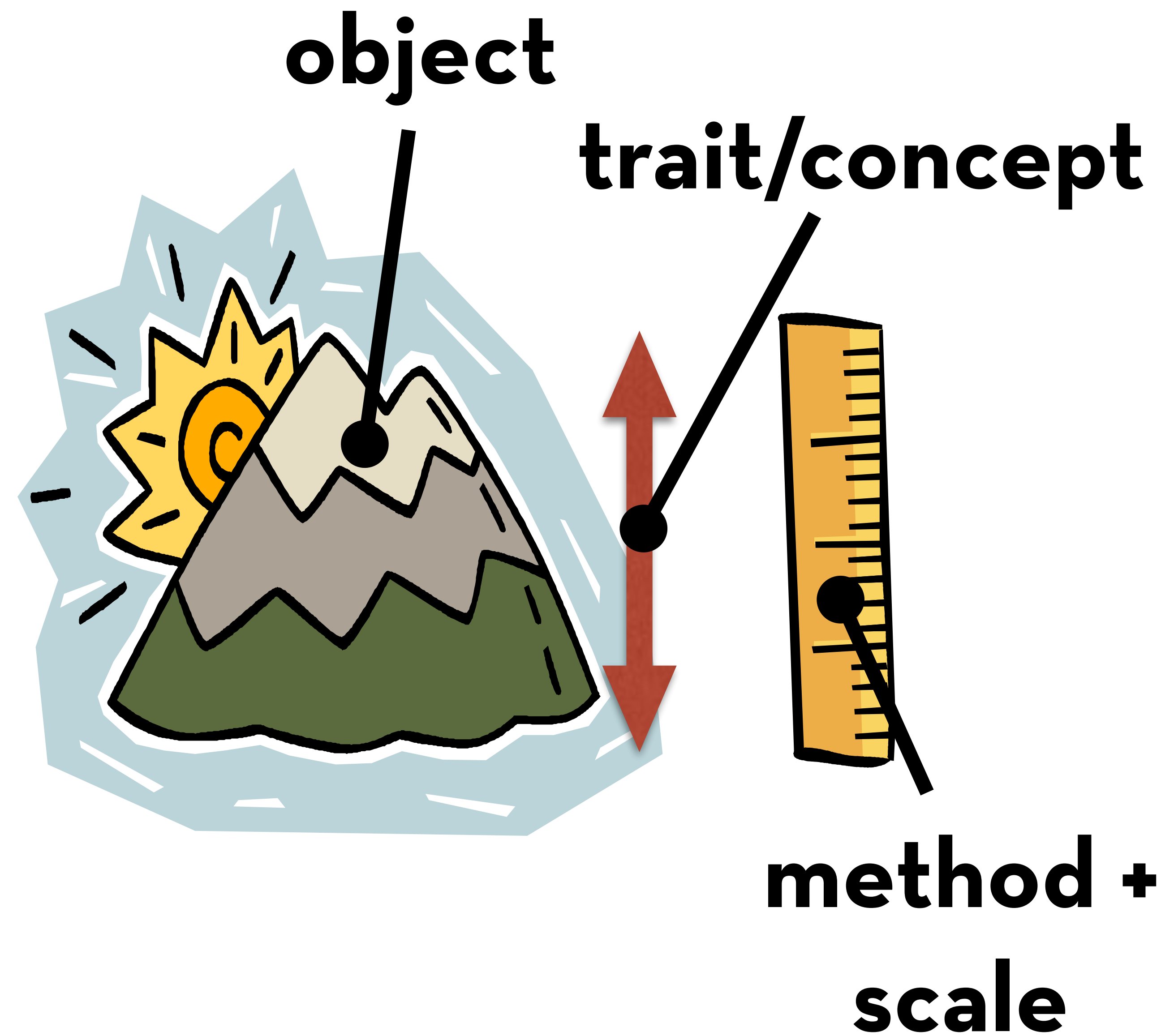
general theory

Measuring things

The quantification of a trait of an object

Using a method

On a 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

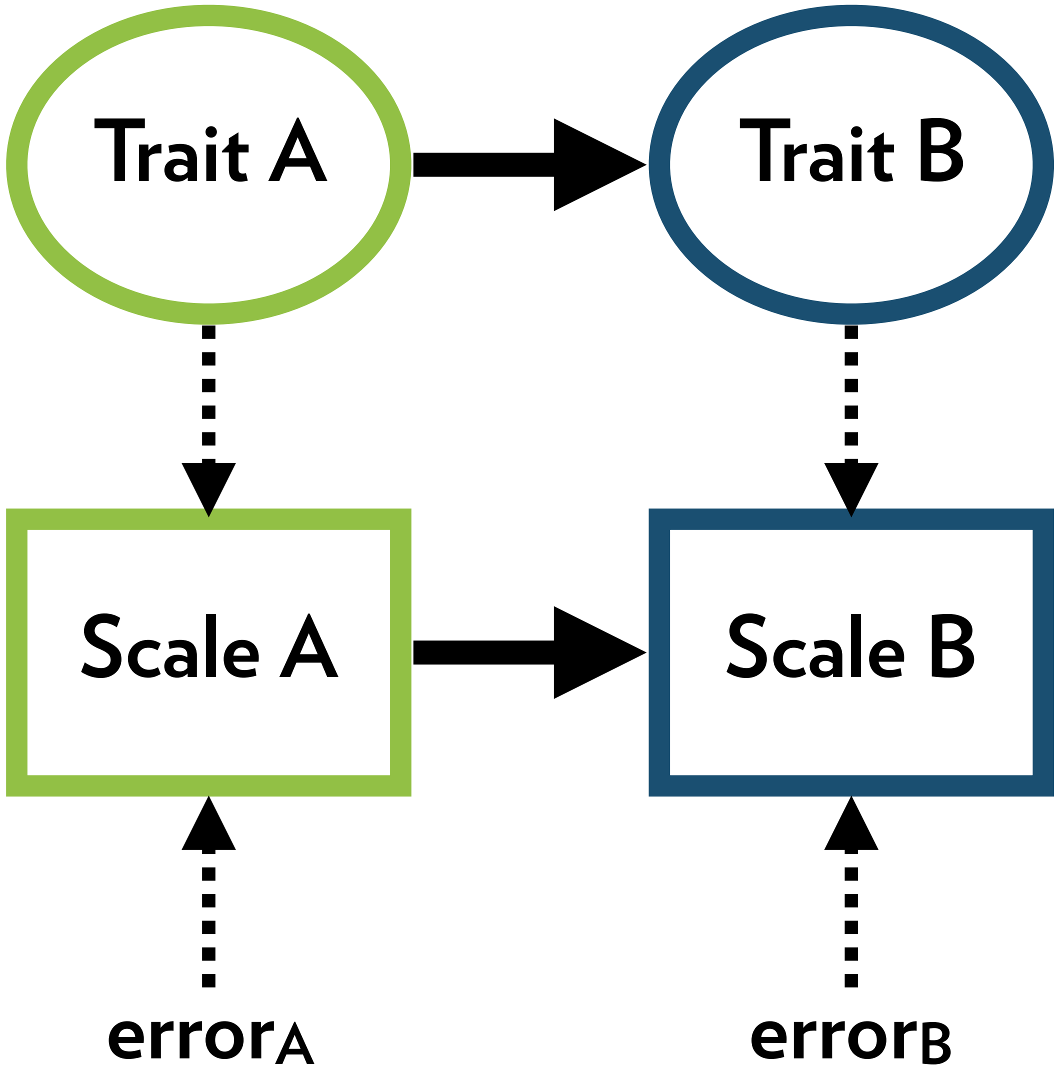
a reason to think about subjective valuations

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



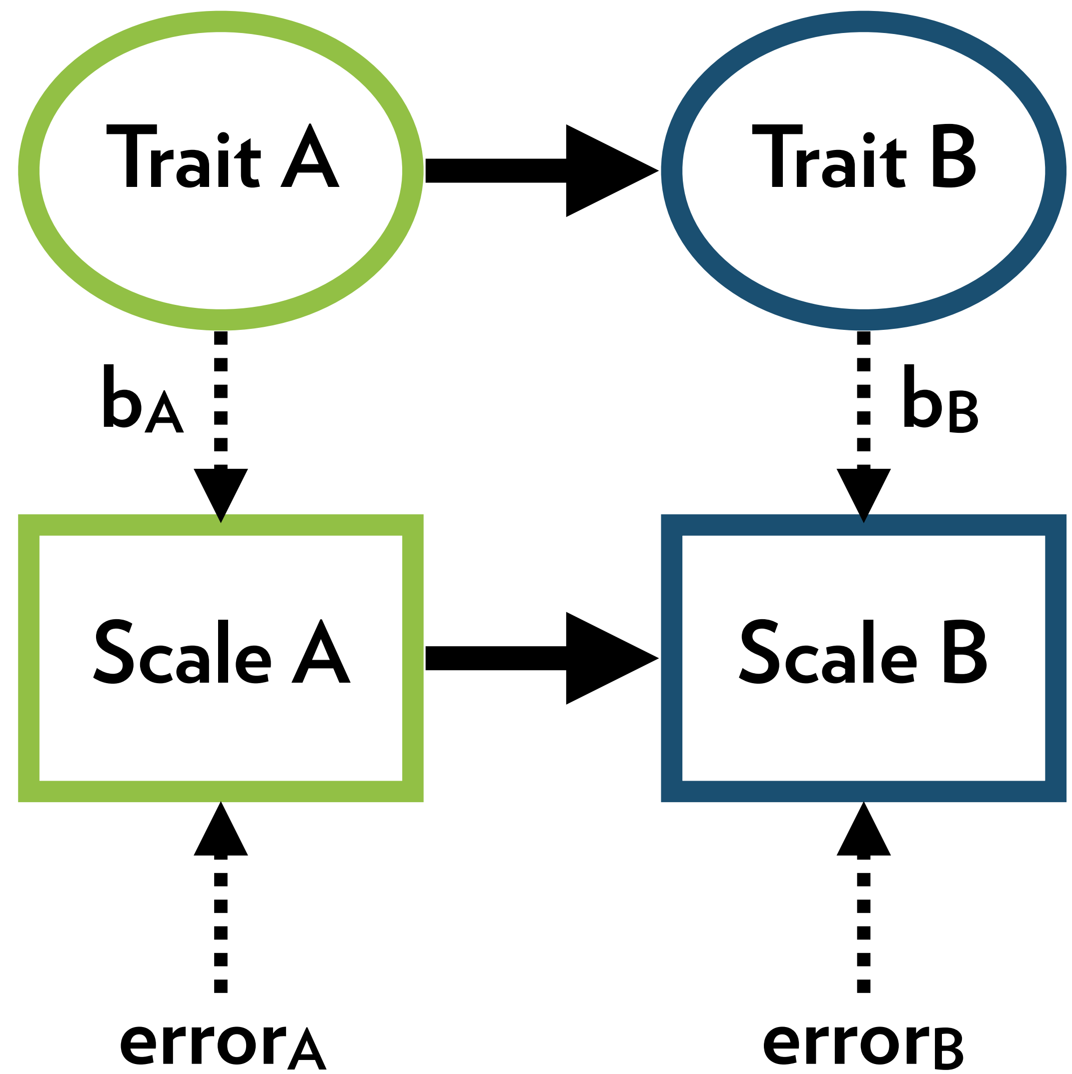
Latent variables

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

$$\text{Scale A} = a + b_A \text{Trait A} + \text{error}_A$$



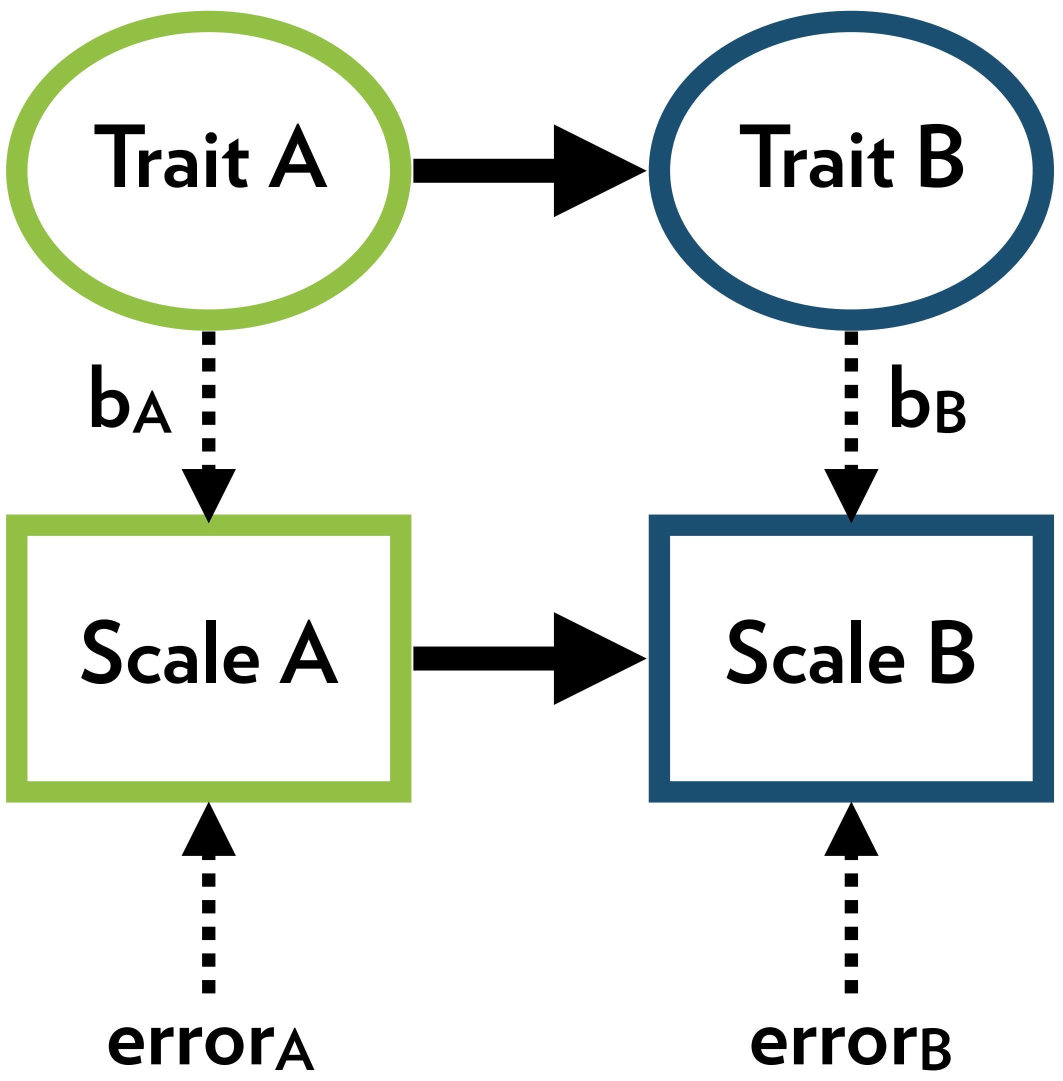
Latent variables

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

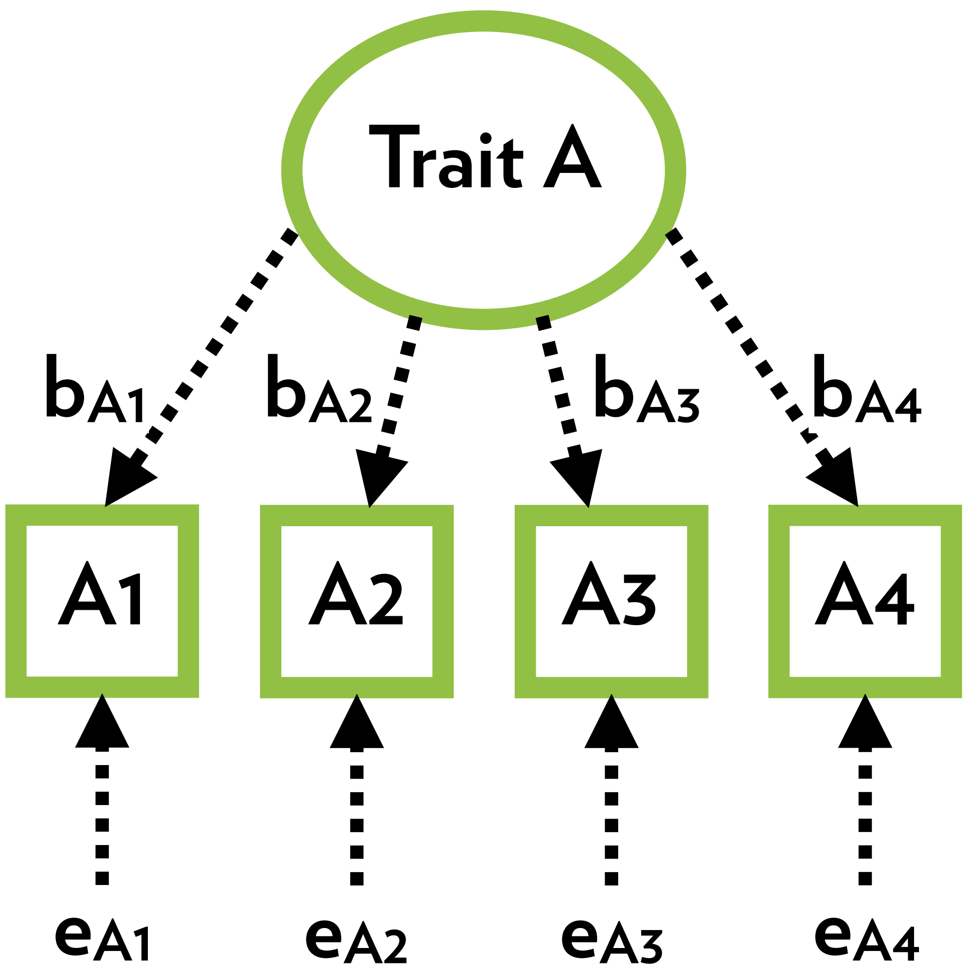
Another reason to have multiple items!



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"

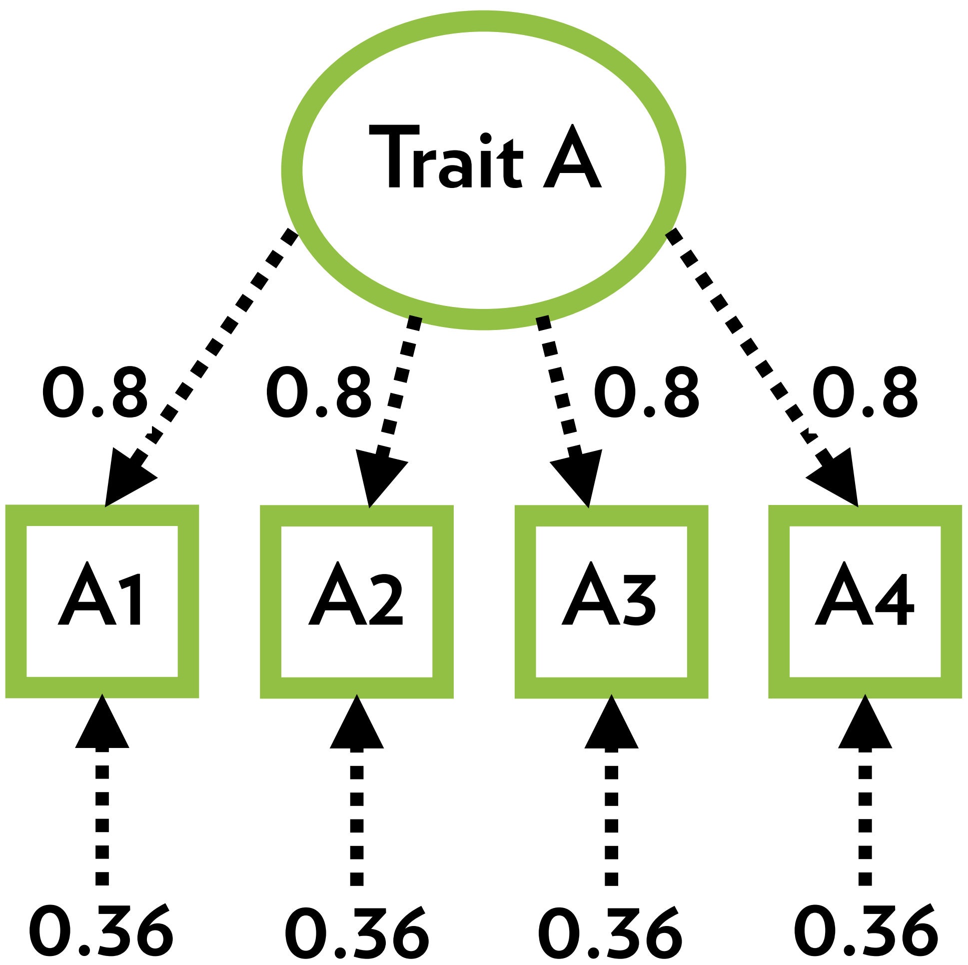


Latent variables

Fill in the numbers:

> To reconstruct the correlations, follow the paths!

(Next week we will do a version of this with multiple traits and unequal correlations)



Reliability

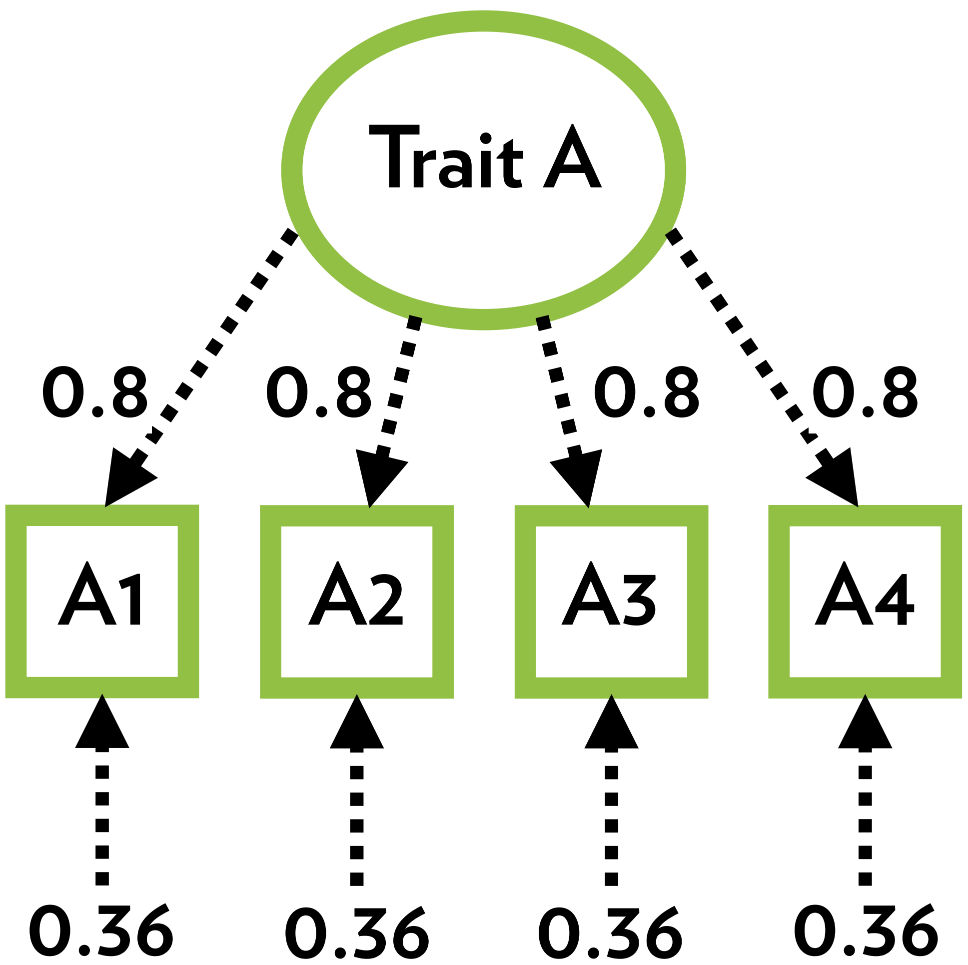
how good is this scale, statistically speaking?

Reliability

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

Consistent scales have:

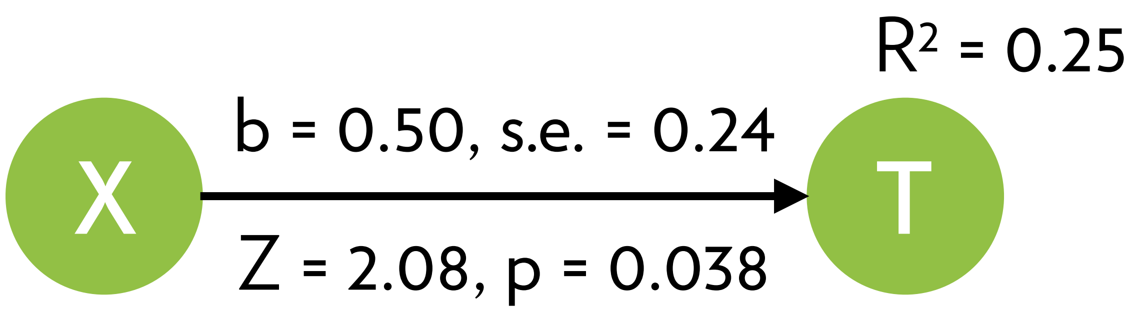
- Low uniquenesses
- High communalities
- High loadings
- High correlation between items



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...

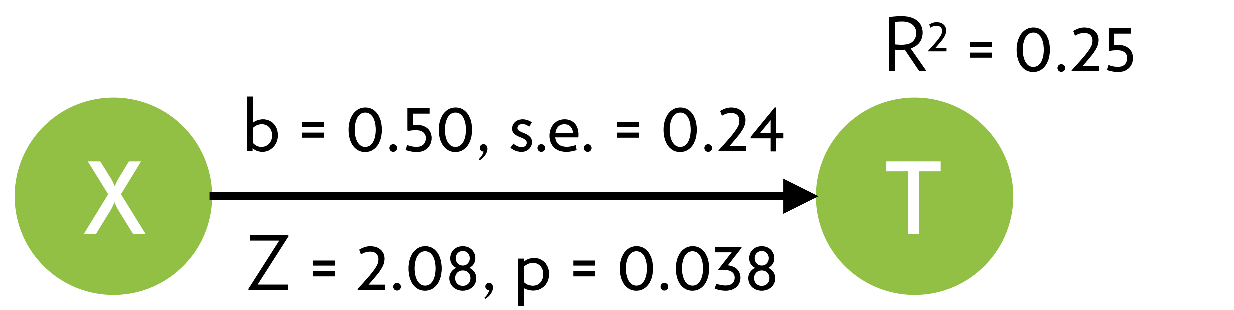


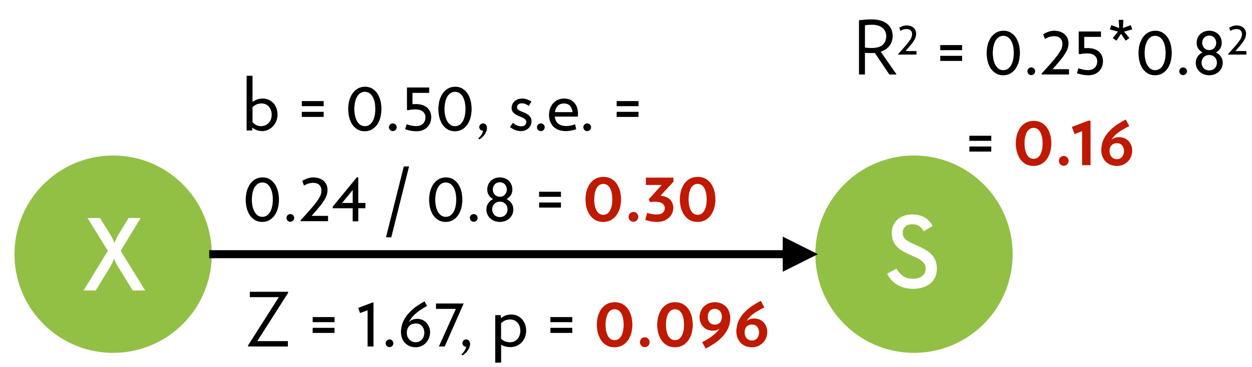
Problems...

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





Solution!

Two weeks from now, we will learn **Structural Equation Modeling**, a method that has 100% power regardless of the reliability of the measurement scales!

Reliability measures

Cronbach's Alpha uses the covariance matrix between items:

alpha = average(Cov) / average(Cov & Var)

Standardized alpha uses the average correlation r:

alpha = kr / (1+(k-1)r), where k is the number of variables

	A	B	C	D
A	Var_A	$Cov_{A,B}$	$Cov_{A,C}$	$Cov_{A,D}$
B	$Cov_{A,B}$	Var_A	$Cov_{B,C}$	$Cov_{B,D}$
C	$Cov_{A,C}$	$Cov_{B,C}$	Var_A	$Cov_{C,D}$
D	$Cov_{A,D}$	$Cov_{B,D}$	$Cov_{C,D}$	Var_A

Reliability measures

Average Variance Extracted (AVE) is the average R^2 of the model

Also: 1-average(e)

Also: average($loading^2$)

This one also works when correlations are unequal!

We will use it next week

Alpha in R

Load twq.dat, variables:

- cgraph: inspectability (0: list, 1: graph)
- citem-cfriend: control (baseline: no control)
- cig (citem * cgraph) and cfg (cfriend * cgraph)
- s1-s7: satisfaction with the system
- q1-q6: perceived recommendation quality
- c1-c5: perceived control
- u1-u5: understandability

Alpha in R

Variables (continued):

- e1-e4: user music expertise
- t1-t6: propensity to trust
- f1-f6: familiarity with recommenders
- average rating of, and number of known items in, the top 10
- time taken to inspect the recommendations

Alpha in R

Use alpha in package "psych":

```
alpha(twq[,c("s1","s2","s3","s4","s5","s6","s7")])
```

Alpha in R

```
  raw_alpha std.alpha G6(smc) average_r S/N  ase mean   sd
      0.92      0.92    0.92      0.64  12 0.02 0.64 0.86

 lower alpha upper     95% confidence boundaries
0.88 0.92 0.96

 Reliability if an item is dropped:
    raw_alpha std.alpha G6(smc) average_r  S/N alpha se
s1       0.91      0.91    0.90      0.62  9.9    0.024
s2-      0.91      0.91    0.90      0.62  9.9    0.024
s3       0.92      0.92    0.91      0.65 11.2    0.024
s4       0.91      0.91    0.90      0.64 10.7    0.024
s5       0.90      0.91    0.90      0.62  9.7    0.025
s6       0.92      0.92    0.91      0.66 11.6    0.023
s7-      0.91      0.91    0.90      0.64 10.4    0.024

 Item statistics
      n raw.r std.r r.cor r.drop mean   sd
s1  267  0.86  0.86  0.84   0.81 0.67 1.02
s2- 267  0.86  0.86  0.84   0.81 0.99 1.04
s3  267  0.79  0.79  0.74   0.72 0.46 1.03
s4  267  0.82  0.82  0.78   0.74 0.38 1.09
s5  267  0.88  0.87  0.85   0.82 0.41 1.08
s6  267  0.75  0.77  0.71   0.68 1.10 0.79
s7- 267  0.84  0.83  0.80   0.77 0.43 1.21
```

Alpha in R

Output includes:

- raw_alpha: Chronbach's Alpha
- std.alpha: Standardized Alpha
- average correlation between items
- The values of these metrics if any item is dropped
- raw.r: correlation of item with scale
- cor.r: partial correlation of item with scale, adjusted for reliability
- drop.r: correlation of item with scale without the item

Validity

how good is this scale, practically speaking?

Validity

Reliability: How well does the scale measure the latent variable?

Validity: Is the latent variable really the thing we wanted to measure?

Note: validity is always assessed in **context**! It depends on:

- the specific **population** to be measured
- the **purpose** of the measure

Types of validity

Content validity (face validity)

Criterion validity

- Predictive validity
- Concurrent validity

Construct validity

- Discriminant validity
- Convergent validity

Content validity

Content validity is assessed by specialists in the concept to be measured

Do the items cover the breath of the content area? (not too wide, not too narrow?)

Are they in an appropriate format?

Bad:

- A attitude scale that also has behavioral items
- A usability scale that only asks about learnability
- A relative measure of risk, trying to measure absolute risk

Criterion validity

Predictive validity

Test how well a measure predicts a future outcome (e.g. behavioral intention —> future behavior)

Concurrent validity

Compare the measure with some other measure that is known to correlate with the concept (e.g. correlate a new scale for altruism with an existing scale for compassion)

Or, compare the measure between groups that are known to differ on the concept (e.g. compare altruism of nuns and homicidal maniacs)

Construct validity

Discriminant validity

- Are two scales really measuring different things? (e.g. attitude and satisfaction may be too highly correlated)

Convergent validity (= reliability)

- Is the scale really measuring a single thing? (e.g. a usability scale may actually consist of several sub-scales: learnability, effectiveness, efficiency, satisfaction, etc.)

Factor analysis gives you both types of construct validity

- Other types you have to confirm yourself!

"It is the mark of a truly intelligent person to be moved by statistics."

T H A N K S !

George Bernard Shaw