Workshop “Test Construction and Test Analysis with R”

Wolfgang Lenhard[1] & Boromir, Son of Denethor

[1]Institute of Psychology

Universität Würzburg


ONE DOES NOT SIMPLY ...
... LEARN PSYCHOMETRICS
BY LISTENING TO A TALK ...
imgflip.com

Schedule

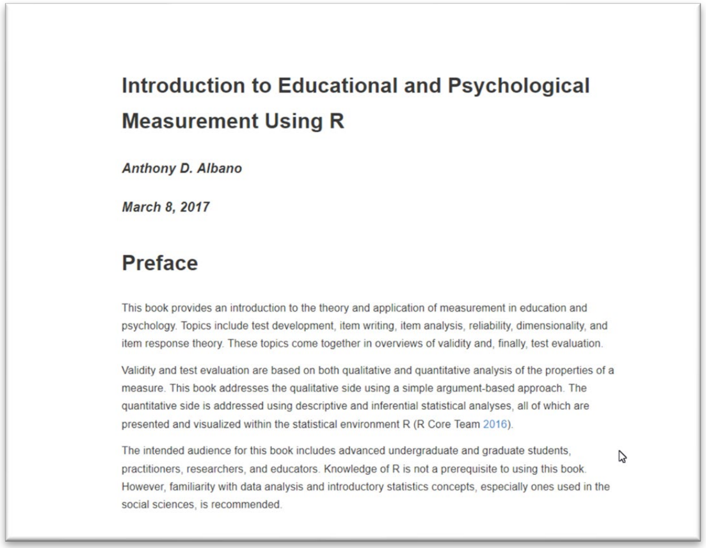
Day 1

1. A very brief introduction to test construction: The basics of Classical Test Theory (CTT) and Item Response Theory (IRT)
2. A very brief introduction to R: The environment, installing packages, basic data structures (objects, vectors & data frames), functions

Day 2

1. Item analysis (difficulty, discrimination, differential item functioning)
2. Scale analysis (homogeneity & reliability, unidimensionality, IRT scaling, specific objectivity / model tests)
3. Norming
4. Measurement invariance

Some literature hints …



Introduction to Educational and Psychological Measurement Using R

Anthony D. Albano

March 8, 2017

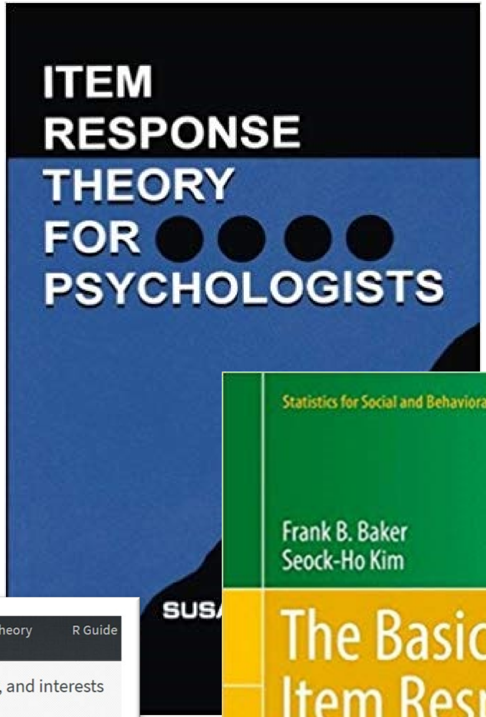
Preface

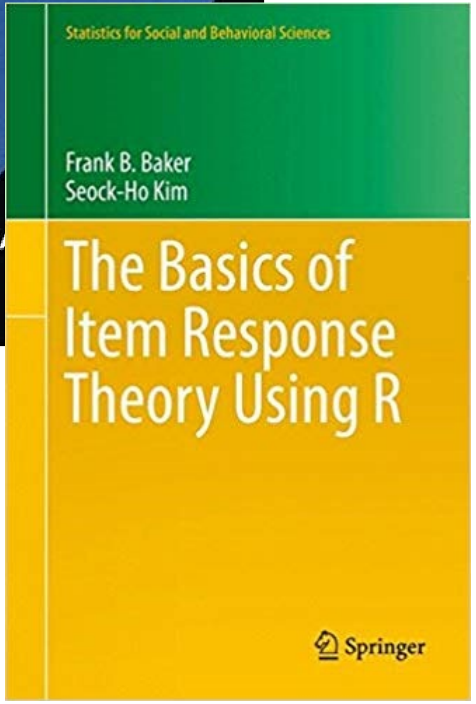
This book provides an introduction to the theory and application of measurement in education and psychology. Topics include test development, item writing, item analysis, reliability, dimensionality, and item response theory. These topics come together in overviews of validity and, finally, test evaluation.

Validity and test evaluation are based on both qualitative and quantitative analysis of the properties of a measure. This book addresses the qualitative side using a simple argument-based approach. The quantitative side is addressed using descriptive and inferential statistical analyses, all of which are presented and visualized within the statistical environment R (R Core Team 2016).

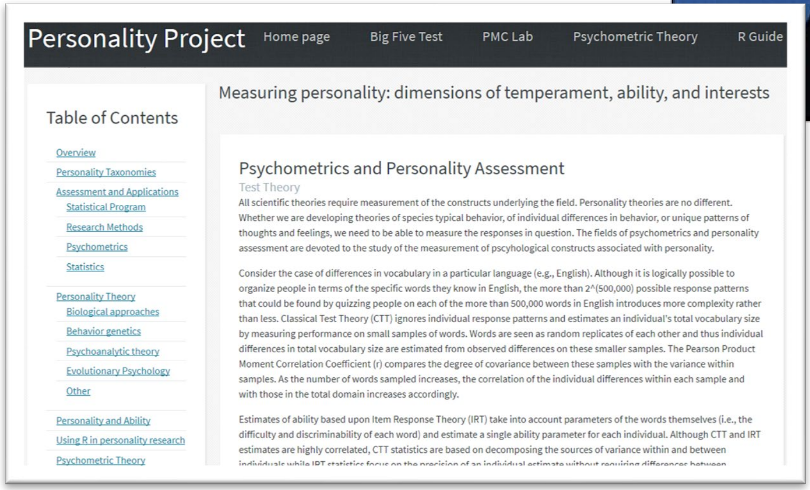
The intended audience for this book includes advanced undergraduate and graduate students, practitioners, researchers, and educators. Knowledge of R is not a prerequisite to using this book. However, familiarity with data analysis and introductory statistics concepts, especially ones used in the social sciences, is recommended.

https://www.thetaminusb.com/intro-measurement-r/





https://cran.r-project.org/



Personality Project | Home page | Big Five Test | PMC Lab | Psychometric Theory | R Guide

Measuring personality: dimensions of temperament, ability, and interests

Table of Contents

- Overview
- Personality Taxonomies
- Assessment and Applications
  - Statistical Program
  - Research Methods
  - Psychometrics
  - Statistics
- Personality Theory
  - Biological approaches
  - Behavior genetics
  - Psychoanalytic theory
  - Evolutionary Psychology
  - Other
- Personality and Ability
- Using R in personality research
- Psychometric Theory

Psychometrics and Personality Assessment

Test Theory

All scientific theories require measurement of the constructs underlying the field. Personality theories are no different. Whether we are developing theories of species typical behavior, of individual differences in behavior, or unique patterns of thoughts and feelings, we need to be able to measure the responses in question. The fields of psychometrics and personality assessment are devoted to the study of the measurement of pscyhological constructs associated with personality.

Consider the case of differences in vocabulary in a particular language (e.g., English). Although it is logically possible to organize people in terms of the specific words they know in English, the more than 2^(500,000) possible response patterns that could be found by quizzing people on each of the more than 500,000 words in English introduces more complexity rather than less. Classical Test Theory (CTT) ignores individual response patterns and estimates an individual's total vocabulary size by measuring performance on small samples of words. Words are seen as random replicates of each other and thus individual differences in total vocabulary size are estimated from observed differences on these smaller samples. The Pearson Product Moment Correlation Coefficient (r) compares the degree of covariance between these samples with the variance within samples. As the number of words sampled increases, the correlation of the individual differences within each sample and with those in the total domain increases accordingly.

Estimates of ability based upon Item Response Theory (IRT) take into account parameters of the words themselves (i.e., the difficulty and discriminability of each word) and estimate a single ability parameter for each individual. Although CTT and IRT estimates are highly correlated, CTT statistics are based on decomposing the sources of variance within and between

http://personality-project.org/

Test Construction and Test Analysis with R

VERY VERY BRIEF BASICS

Definitions

- **Test**
  Collection of items intended to measure one or more abilities of a person

- **Item**
  Smallest rateable part of a test (either dichotomous ‘wrong / true’, ordered or interval level; in case of performance assessments it is dichotomous in most cases)

- **Scale**
  Aggregate of items that measure the same ability

- **Latent ability**
  Trait / characteristic of a person that is not directly observable (e. g. intelligence, reading comprehension, vocabulary knowledge …) and which has to be indirectly inferred from test results

- **Measurement**
  “Homologous transformation of an empirical relative into a numerical relative”: Assessment of the rank order, the magnitude of a characteristic, the relative distances between different people... so that the real situation is adequately expressed by numbers. The relation and magnitude of features within a sample of persons has to be preserved

Method inventory: "Classical test theory (CTT)"

- Classical test theory aims at quantifying errors of measurement:

$$Observed\ Score = True\ Score + Error$$

(Precondition: Errors are not correlated and distributed normally)

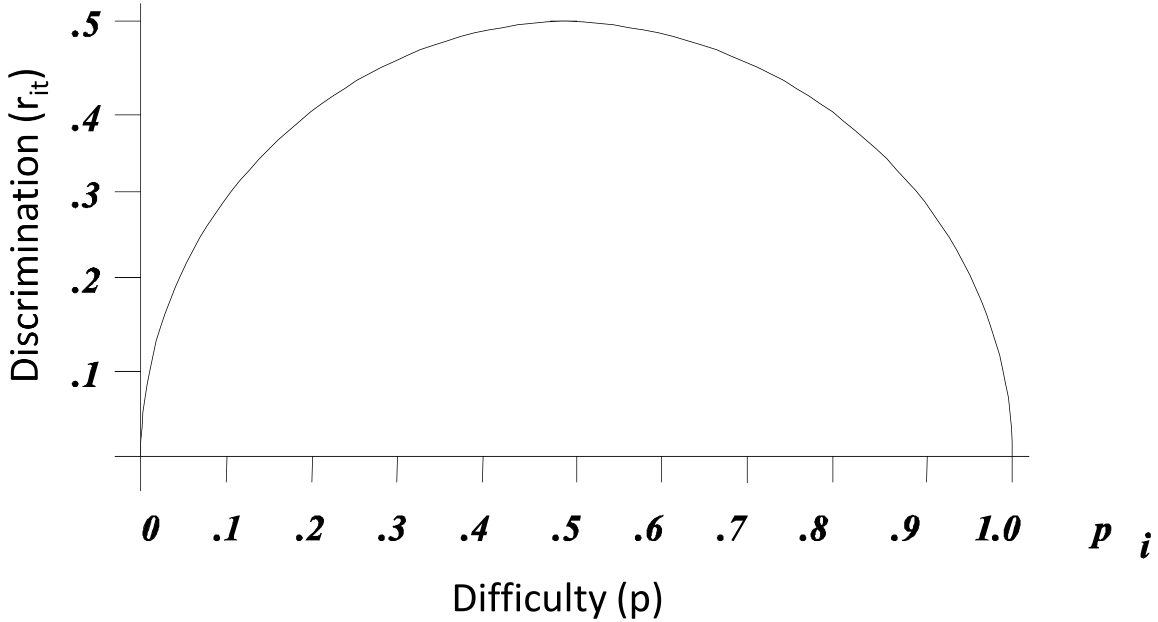
- Ability = Raw Value of Scale ± Confidence Interval
- Reliability:

$$Reliability = \frac{Total\ Variance\ - Error\ Variance}{Total\ Variance} = 1\ - \frac{Error\ Variance}{Total\ Variance}$$

- Measures of Reliability: retest correlation r_{tt}, split-half, homogeneity α, ω, ICC …
- Difficulty ("Easiness") = Probability of Success
  - 0: Item is not solvable ("floor effect")
  - 1: Every person solves item ("ceiling effect")
  - Guessing probability in case of multiple choice items: $p\ = \frac{1}{k}$

Relationship between item discrimination and difficulty

- Discrimination (r_{it}): Correlation of each single item with total score (minus single item)
  - ⇨ How well does the single item represent / predict the total value?
  - ⇨ Look out for values close to 0; in any case delete negative values
    By convention, values above $r_{it} > .2$ or $r_{it} > .3$ indicate a good fit



- Items with medium difficulty have the highest degree of information / "diagnosticity"
- These items differentiate best between persons who solve an item and persons who do not solve an item.
- Please consider guessing probability:
  MC-item, $k = 4$: $p = .25 + .75 / 2 = .625$

Drawback of CTT:

- Discrimination not in dependence of ability level
- No ability level confidence intervals / measurement error
- Purely data driven (no relation to latent trait)

Item-Response-Theoy by the example of 1PL-Model

- Item-Response-Theory: Calculation of probability of success based on the estimated latent ability:

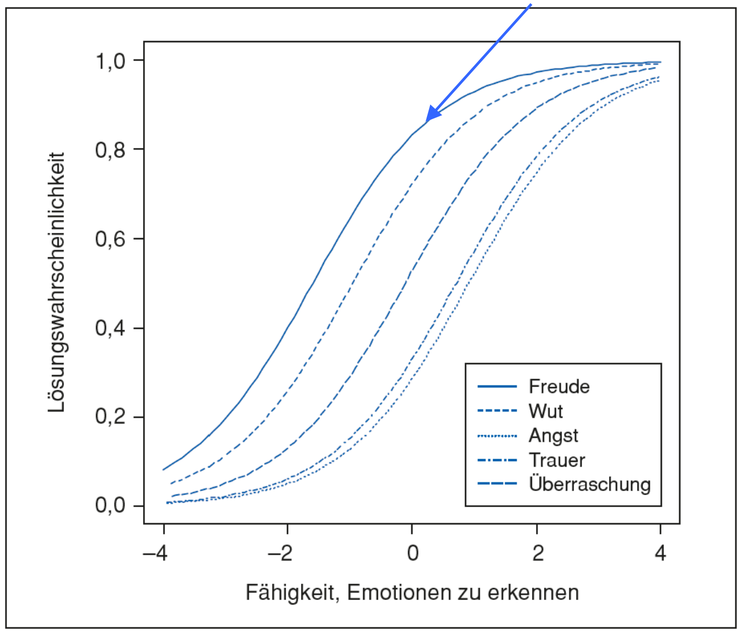
$$P(Y_i = 1|\theta) = \frac{e^{\theta-\alpha_i}}{1+e^{\theta-\alpha_i}}$$

θ = Ability of the person

α_i = Difficulty of item i

- Items differentiate best, when ability and difficulty have the same magnitude
- Difficulty and ability are measured on the same scale
- Values: „logits“ (no fixed range of values; usually centring around 0)
- Drawbacks: Hard to include reaction times / tests with cutoffs (⇨ package `LNIRT`)

Item Characteristics Curve (ICC)

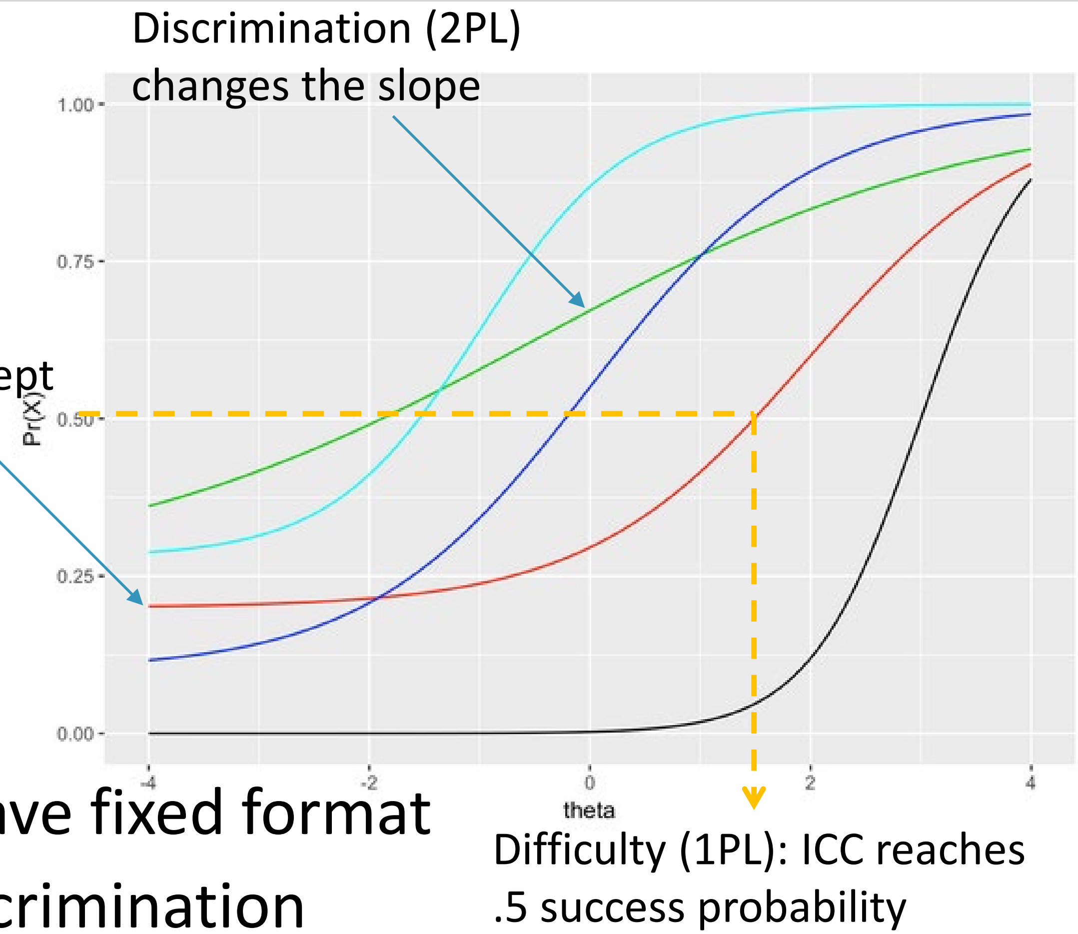


Eid & Schmidt, 2010

What does 1PL, 2PL, 3PL mean?

- ICCs can be modified by a discrimination index (= 2PL) and a guessing parameter (= 3 PL)
  a. Difficulty
  b. Discrimination
  c. Guessing

$$P(Y_i = 1|\theta) = c_i + (1 - c_i) \cdot \frac{e^{b_i(\theta-\alpha_i)}}{1 + e^{b_i(\theta-\alpha_i)}}$$



- Rasch is a 1PL model, Birnbaum a 2 PL
- Guessing is unnecessary to include, if items have fixed format
- 1 PL assumes that all items have the same discrimination (very strong assumption) and Rasch adds further model assumptions (specific objectivity …)

IRT vs. CTT: What should we use?

There is no contradiction:

- Both give you valuable methods which complement each other.
- You can apply both in the same analysis depending on what you want to know.
- IRT is much more powerful, but needs larger datasets. CTT focuses on scale level.
- Additional analyses and features of IRT are not always necessary.
- Often both methods identify the same problematic items.

“If you only have a hammer, every problem looks like a nail”, or …

Use, what fits your research question best. Combine different methods according to your needs! To this ends, you need the according rich method inventory.

So, what levels of analysis are generally used?

- Single items:
  - Difficulty (in CTT it rather should be called ‘easiness’)
  - Discrimination (part-whole correlation with total scale value)
  - Fairness („Differential Item Functioning“)
  - In general: diagnostic information of single items
  - …
- Scale:
  - Homogeneity
  - Dimensionality
  - Reliability
  - Validity of model assumptions (model fit, specific homogeneity, local stochastic independence …)
  - …
  - Measurement invariance
  - Norm data modelling

Test Construction and Test Analysis with R

STEPS IN TEST CONSTRUCTION

Basic steps in test construction

Planing stage

1. Theoretical derivation
2. Collect / generate items
3. Complete first draft of the test

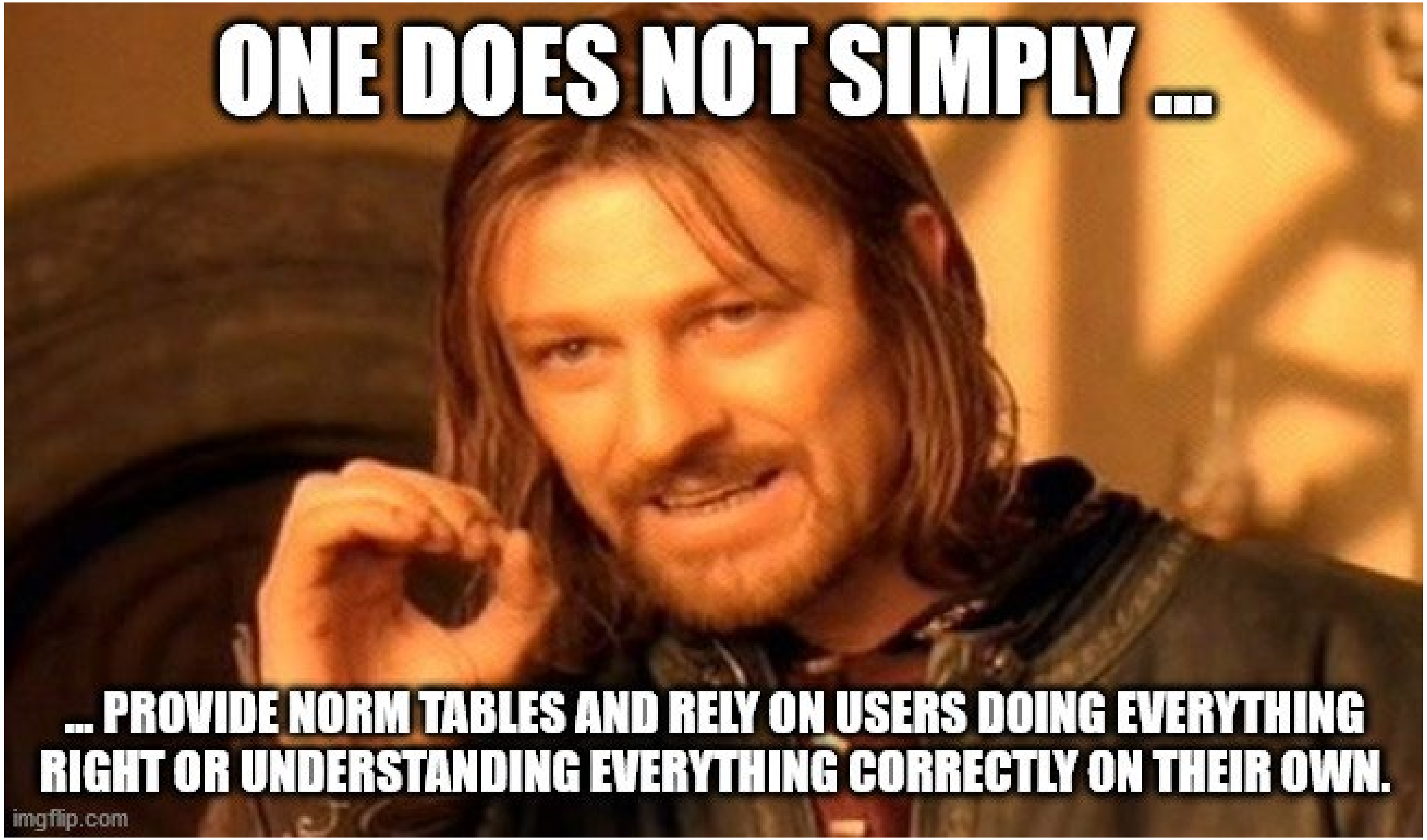
Evaluation stage

4. Pilot study
5. Item analysis and selection ⇨ Final test version
6. Standardization
7. Assessment of quality criteria, test analysis

Norming, scaling, equating

8. Collection of a representative norming sample
9. Stratification
10. Norm score modelling / linking & equating / adaptive testing strategies …
11. Documentation (Sidenote: NEVER report z scores, don’t have users add up positive & negative digit numbers)

In case, you plan to publish your test …



Checklist for selecting / generating items

1. **Generating items: “There is nothing more practical than a good theory” (Kurt Lewin)**
   - Generate items that represent your construct
   - Choose an item format and use that stringently throughout your scale: Test construction is not about formal creativity! Choose formats, that can be assessed unambiguously.

2. After pilot study: Check difficulty and distribution
   - Very easy items can be used as “icebreakers”.
   - Very difficult items help to avoid ceiling effects.
   - Variation in item difficulty helps to cover different ability levels.
   - It makes sense to rank/sort the items according to difficulty so as not to frustrate people with test anxiety.
   - Items with difficulty below the guessing probability are questionable: Check distractors!

3. Analyse discrimination and homogeneity
   - Define selection criteria, in any case removal of items with negative discrimination (unless they exhibit good features in specific subgroups, where you need them)
   - In case of item exclusion: Does homogeneity improve?

Checklist for selecting / generating items

3. Testing model assumptions
   - Is scale unidimensional? (several possibilities, e. g. EFA / CFA)
   - Are model assumptions valid (e. g. comparison of model parameters between different subsamples ⇨ specific objectivity; local stochastic independence …)?

4. Check fairness of items: Differential Item Functioning (DIF; uniform; non-uniform), e. g. gender
   - Uniform DIF: $item_i = intercept + \ b_{i1} \cdot scale + \ \boldsymbol{b_{i2} \cdot gender} + e_i$
   - Non-Uniform DIF: $item_i = intercept + \ b_{i1} \cdot scale + \ b_{i2} \cdot gender + \ \boldsymbol{b_{i3} \cdot gender \cdot scale} + e_i$

In general:

- Do not exclude items without thorough evaluation. A DIF may contain an important information.
- You need sufficient sample sizes from pilot studies, if possible n > 100, for IRT n > 250

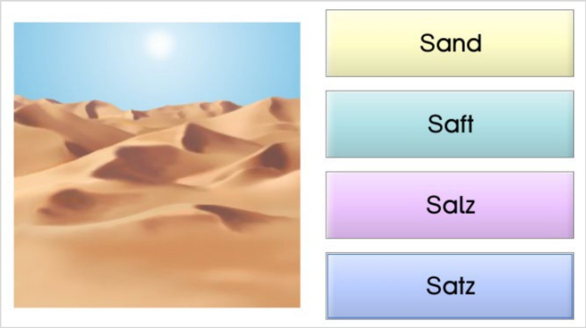
Test Construction and Test Analysis with R

HANDS ON!

Current Study and Test Material

Word Level

- Reading fluency measure
- Low error rate, discrimination of reading ability mainly via speed
- 75 items, 3 minutes time cut off

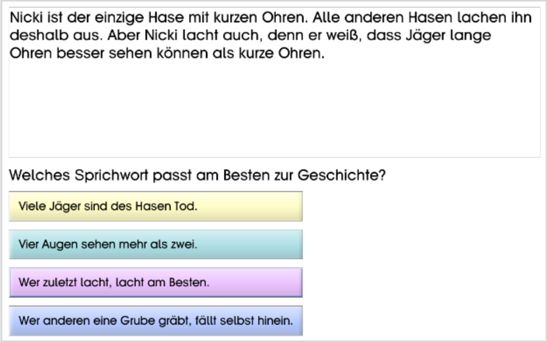


Sentence Level

- Reading fluency and syntactic features
- 36 items, 3 minutes time cut off

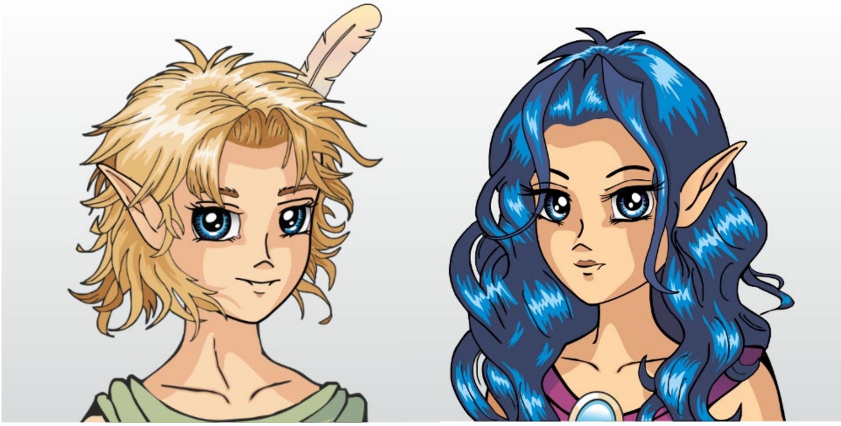
Text Level

- Passage comprehension (gist and verbatim) with narrative and expository texts
- Local and global coherence
- 26 items, 7 minutes time cut off
- For workshop: all items, no cutoff



Data

- Norming project (ELFE-II; W. Lenhard, A. Lenhard & W. Schneider, 2017)
- German norming sample (either computer delivered or paper and pencil)
- Currently a major test of norm based comprehension assessment in Germany (roughly 10 000 copies of the test sold since 2017)

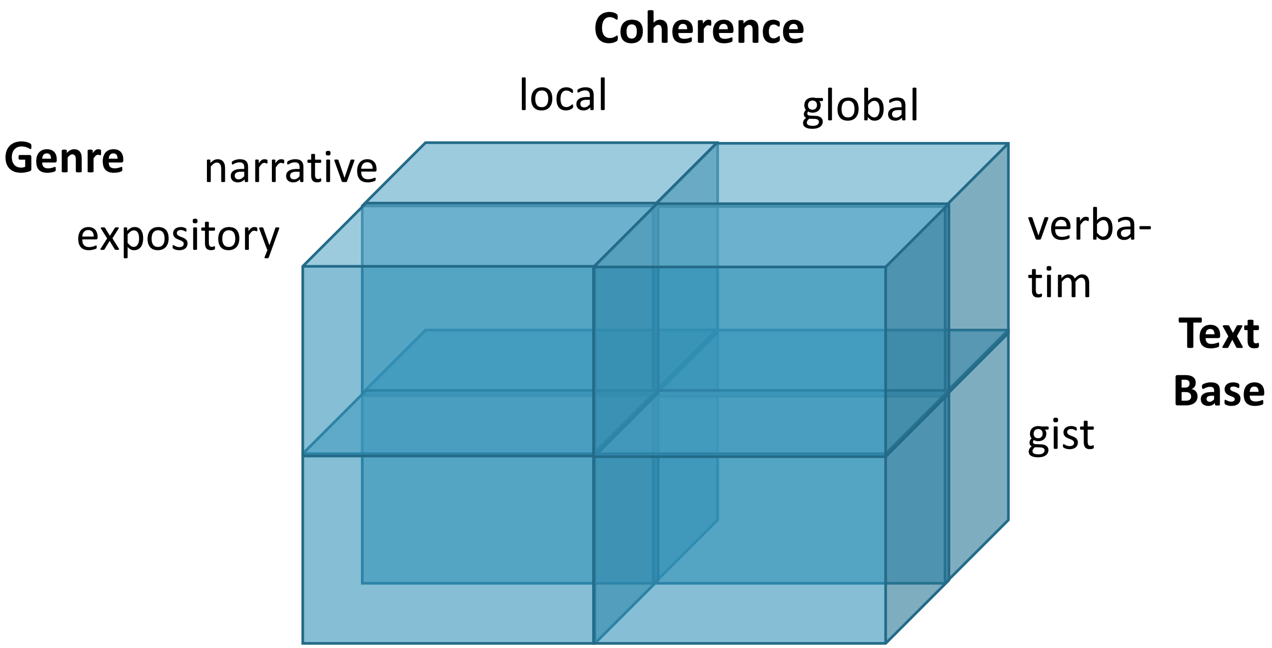


Test Revision	Pilot study (N ~ 620)	Test Revision	Norming	Validation …
June 2014	Sept. & Oct. 2014		May /June 2015	Dec. 2015

- Age range: Grade 1 to 7 (approx. 6 to 14 years)
- Complete sample ≥ 5 000 cases. Stratified norm sample:
  - Paper: n = 1520
  - Computer: n = 1287

Data of the pilot sample on text level

- 31 Multiple Choice Item with 4 alternatives each
- Model:



- Pilot study: 381 cases from the beginning of grade 3 (≅ age 9 years) to 6 (≅ age 13 years)
- Recording of reaction times and alternatives selected; random item order
- In this workshop: Only analysis of accuracy (for reaction time analyses, have a look at `lnirt`)

Planned analyses

STEP 1: Start up! Installation of necessary libraries, getting accustomed with RStudio and R (Handling data etc.)

STEP 2: Read in data (`haven` or just `base` functionality)

STEP 3: Basic descriptive analyses (`psych`)

STEP 4: Analysis of item discrimination, alpha and omega (`psych`)

STEP 5: Assessment of dimensionality, exploratory factor analysis (`psych`)

STEP 6: 1 PL IRT modelling (`TAM`)

STEP 7: Identifying poorly fitting items via ICCs and fit statistics (`TAM`)

STEP 8: Testing model assumptions and differential item functioning (`eRm, difR`)

STEP 9: Norming and norm model evaluation (`cNORM`)

STEP 10: Confirmatory factor analysis and measurement invariance (Multi-Group Confirmatory Factor Analysis – MGCFA; `lavaan, semTools`)

Test Construction and Test Analysis with R

Step 1: Start up! Installation of necessary Packages, getting accustomed with RStudio and R

Learning objectives

- Brief orientation in RStudio
  - Source and R Script
  - Console
  - Environment
  - Code completion
  - Help
- Very basic introduction to R: Environment and some data structures
- Defining objects in R
- Functions
- Packages – and how to load them

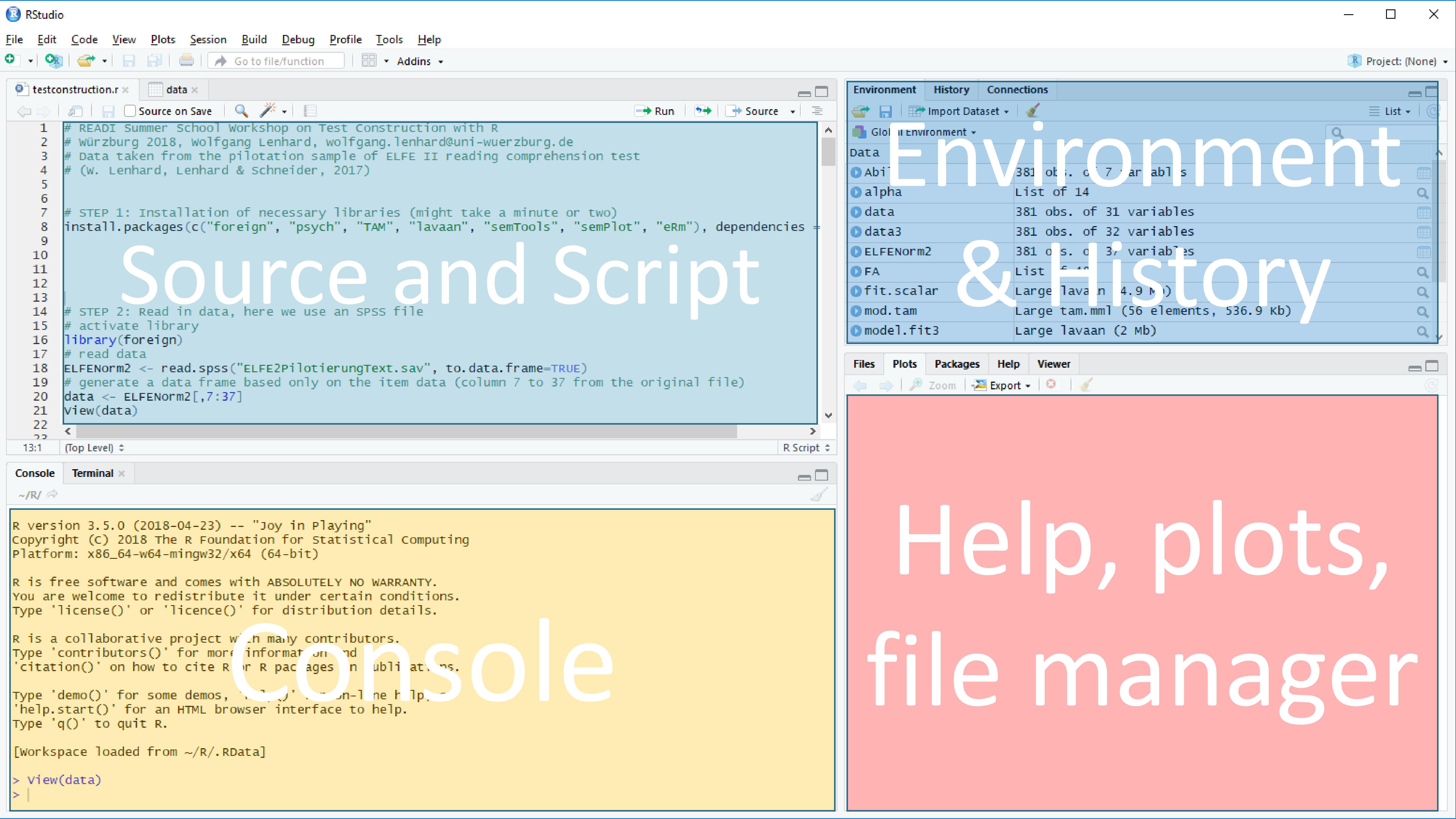
Where to look for help?

- Short intro: https://methodenlehre.github.io/SGSCLM-R-course/
- CrossValidated, Stackoverflow, R-Bloggers …

Basic functions you'll need:

Base, stats and util package (default)

Name	Function
summary, print	Output of results, model summaries, variables; basic functionality of many packages ('S3 functions')
plot, hist (grid, abline, ggf. lm …)	Plotting of results; more advanced packages: ggplot, lattice
View, head, tail	data.frame ist displayed, or head and foot is printed in the console
str	Show structure of objects
load, save	Load and save binary objects (data, functions …)
<-	Assignment operator, e. g. setting up a vector age <- c(8.3, 9.5, 4.2, 5.8)
mean, sd, cor	Basic descriptives
table	Cross table, e. g. frequency per group
?	Display help of function
library	Load a package
data.frame	Data format, most alike to spreadsheet data or SPSS
$	Operator to access a column of a data.frame



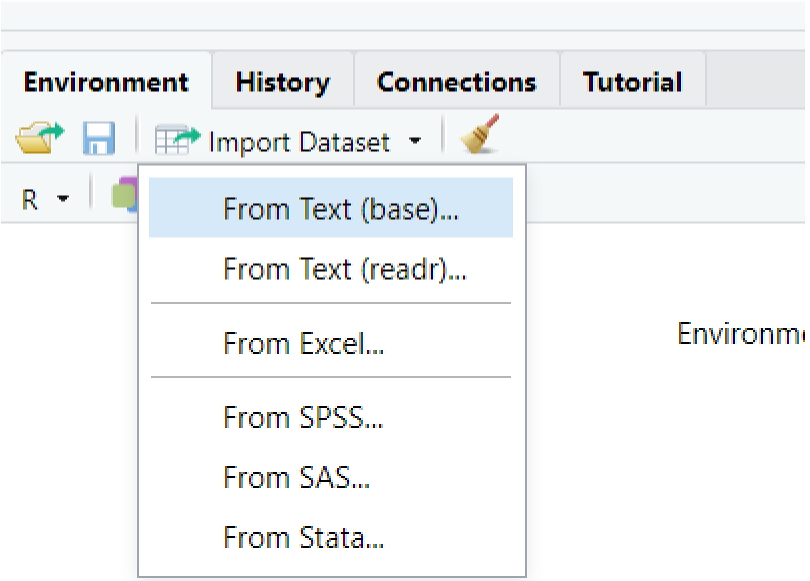
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
testconstruction.r
data
Source on Save
Run
Source
READI Summer School Workshop on Test Construction with R
Würzburg 2018, Wolfgang Lenhard, wolfgang.lenhard@uni-wuerzburg.de
Data taken from the pilotation sample of ELFE II reading comprehension test
(W. Lenhard, Lenhard & Schneider, 2017)
STEP 1: Installation of necessary libraries (might take a minute or two)
install.packages(c("foreign", "psych", "TAM", "lavaan", "semTools", "semPlot", "eRm"), dependencies =
# STEP 2: Read in data, here we use an SPSS file
# activate library
library(foreign)
# read data
ELFENorm2 <- read.spss("ELFE2PilotierungText.sav", to.data.frame=TRUE)
# generate a data frame based only on the item data (column 7 to 37 from the original file)
data <- ELFENorm2[,7:37]
View(data)
Source and Script
Environment History Connections
Import Dataset
List
Data
alpha List of 14
data 381 obs. of 31 variables
data3 381 obs. of 32 variables
mod.tam Large tam.mml (56 elements, 536.9 Kb)
model.fit3 Large lavaan (2 Mb)
Environment & History
Console Terminal
R version 3.5.0 (2018-04-23) -- "Joy in Playing"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/R/.RData]
> View(data)
Console
Files Plots Packages Help Viewer
Zoom
Export
Help, plots, file manager

Test Construction and Test Analysis with R

Step 2: Read in data and try basic data handling

Learning objectives

- Get the data into the environment
- Basic operations with data
- Data handling



Where to look for help?

- Short intros:
  - https://www.thetaminusb.com/intro-measurement-r/
  - https://methodenlehre.github.io/SGSCLM-R-course/
- R for Data Science (Hadley Wickham)

Test Construction and Test Analysis with R

STEP 3: BASIC DESCRIPTIVE ANALYSES

Learning objectives

- describe and describeBy from the ‚psych‘ package
- Plotting means and sd from items

Where to look for help?

- Short intro: https://methodenlehre.github.io/SGSCLM-R-course/
- R for Data Science (Hadley Wickham)

Basic functions you'll need:
psych package

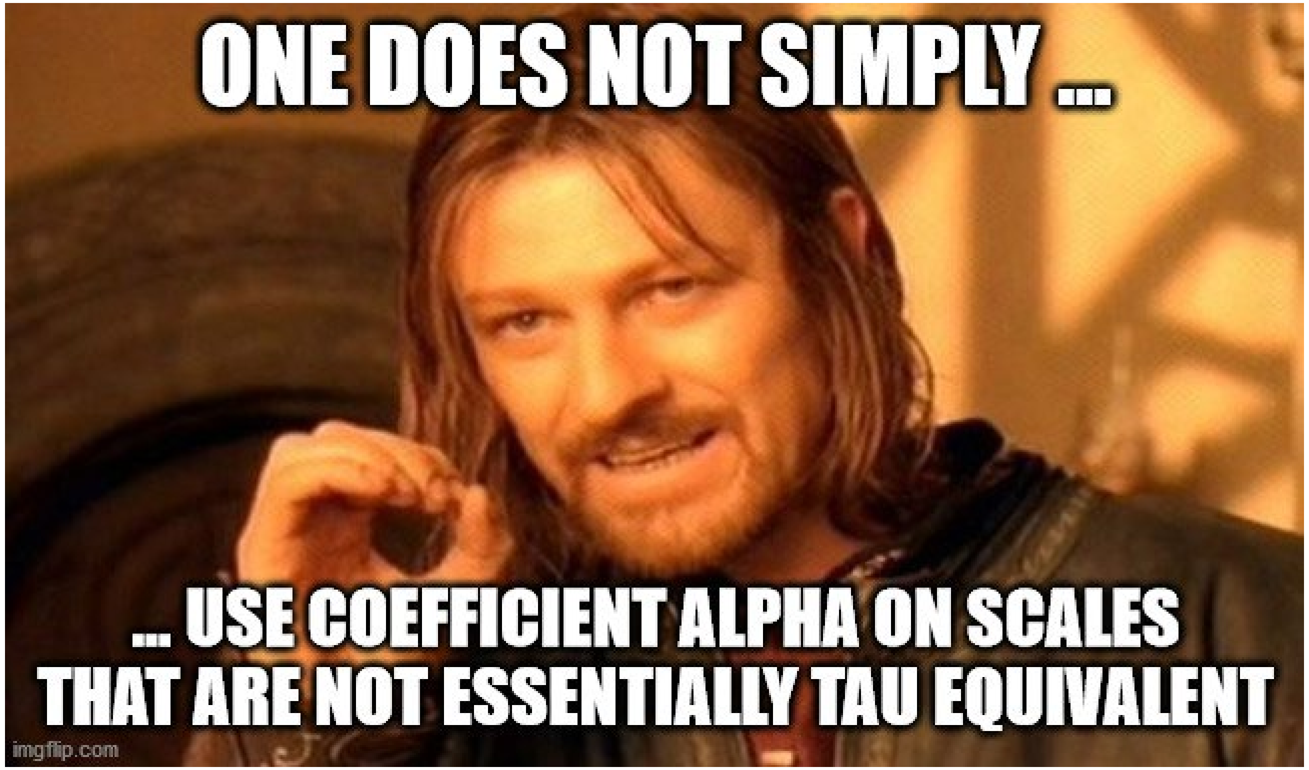
Name	Function
describe, describeBy	Descriptive Statistics
alpha, omega, splitHalf, glb	Homogeneity
fa, fa.diagram, fa.parallel	Factor analyses
TAM and eRm	
tam, tam.wle, tam.mml.2pl, tam.fit	IRT analyses (TAM)
LRtest, plotGOF, Waldtest	Model tests (eRm)

Test Construction and Test Analysis with R

STEP 4: ANALYSIS OF ITEM DISCRIMINATION, ALPHA AND OMEGA

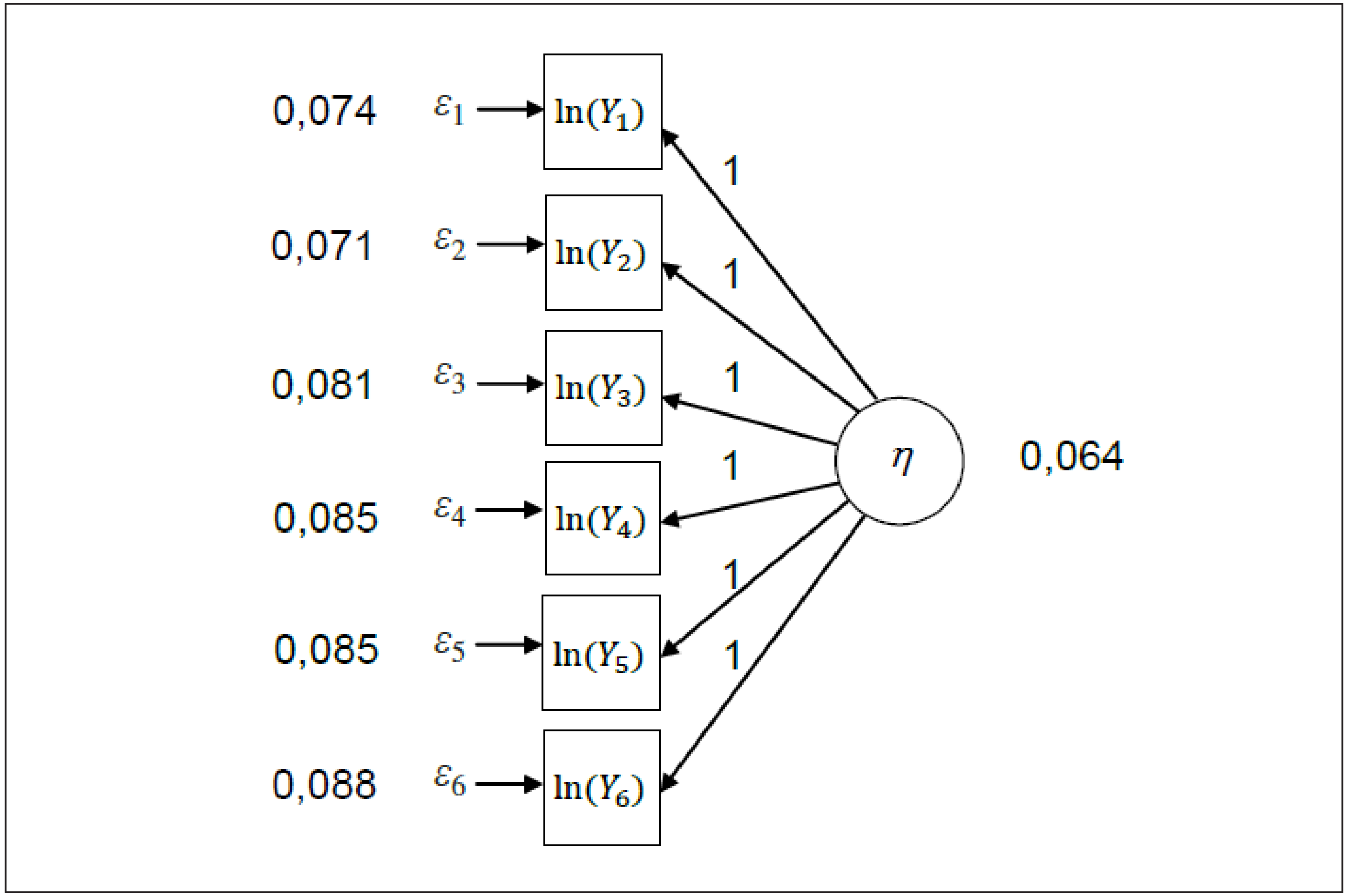
From α to ω

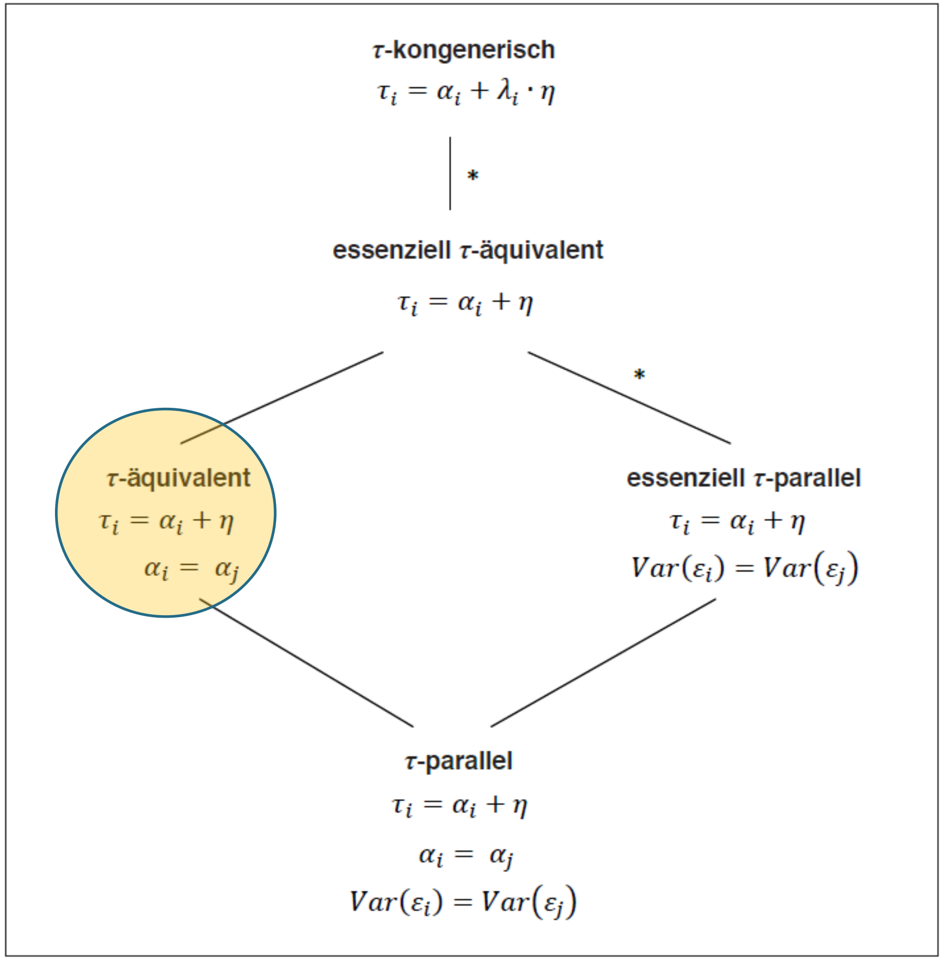
- **Learning objectives:**
  Item selection via discrimation and homogeneity indices



- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144
- Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395–1411. https://doi.org/10.1037/pas0000754
- Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. *Frontiers in Psychology*, *7*, 769. https://doi.org/10.3389/fpsyg.2016.00769

Essential τ equivalence (Eid & Schmidt, 2010, chapter 6)



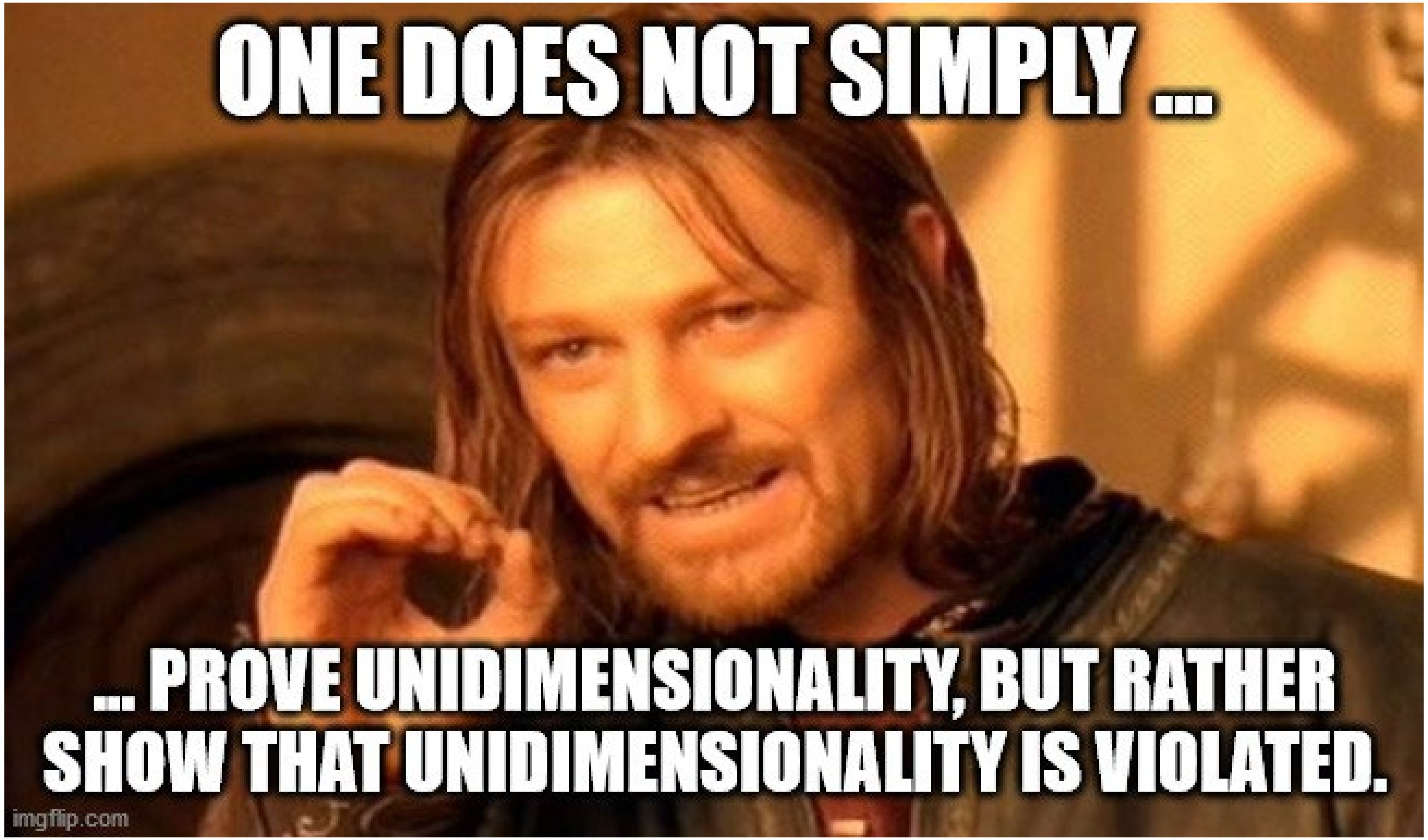


Eid & Schmidt, 2010

Test Construction and Test Analysis with R

STEP 5: ASSESSMENT OF DIMENSIONALITY (PSYCH)

One dimension to rule them all!



Test Construction and Test Analysis with R

STEP 6: 1 PL IRT MODELLING (TAM)

STEP 7: IDENTIFY POORLY FITTING ITEMS VIA ICCS AND FIT STATISTICS PLOT MODEL

Basic functions you'll need:
TAM package

Learning objectives:

- 1 PL IRT modelling
- Identifying poorly fitting items (and persons)

Name	Function
TAM and eRm	
tam, tam.wle, tam.mml.2pl	IRT analyses
tam.fit	Item fit indices

Where to look for help:

- https://www.edmeasurementsurveys.com/TAM/Tutorials/

Test Construction and Test Analysis with R

STEP 8: MODEL TESTS AND DIFFERENTIAL ITEM FUNCTIONING (DIF)

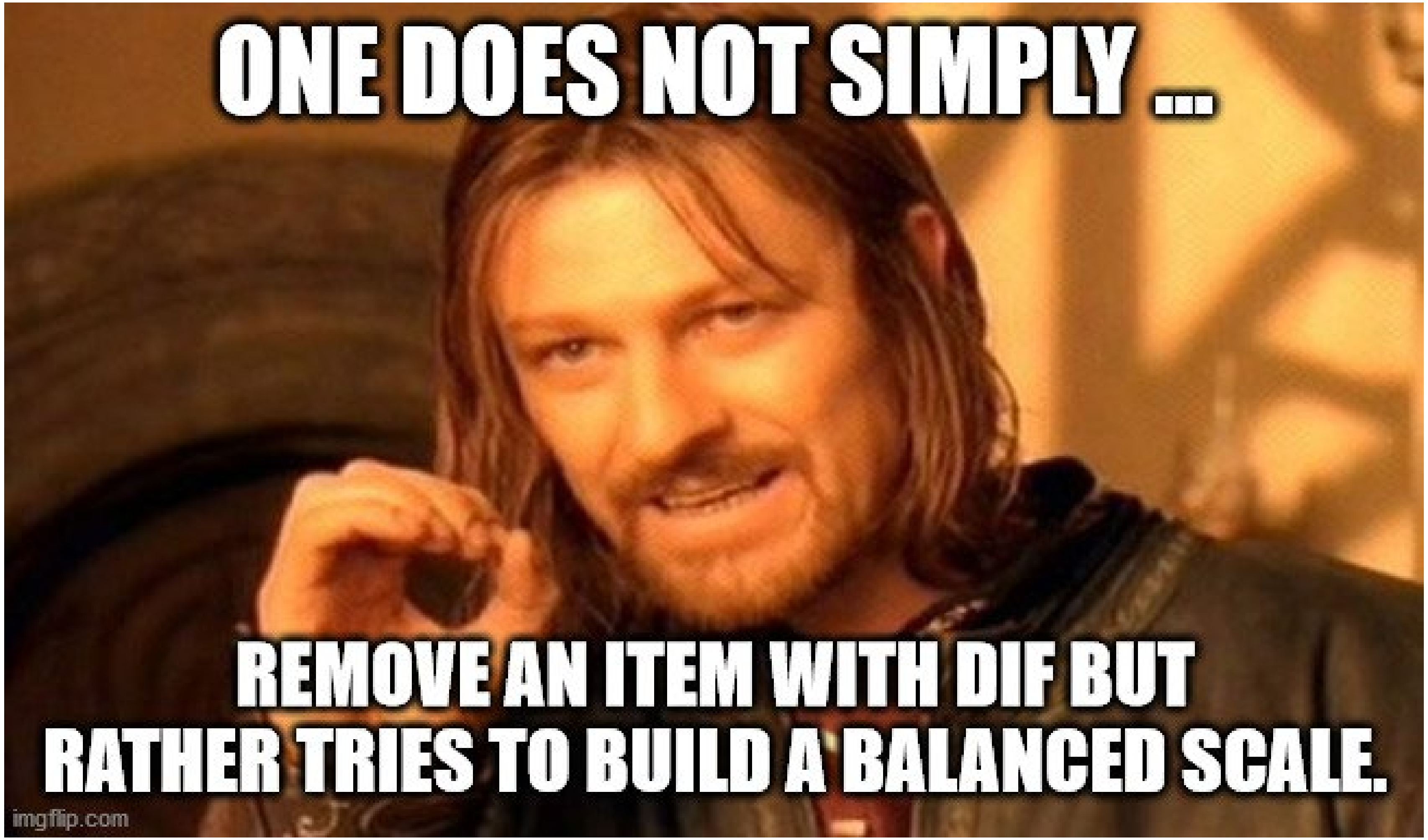
Learning objectives

- Graphical model tests
- Wald test
- Andersen's Likelihood-Ratio Test
- DIF via logistic regression
- Local stochastic independence, multidimensionality and learning

Where to look for help?

- https://hansjoerg.me/2018/04/23/rasch-in-r-tutorial/

You will always find poorly fitting items, when you search long enough …



Basic functions you'll need:
eRm and difR package

Name (eRm)	Function
RM	IRT 1PL model
plotjointICC	Plot ICCs
Itemfit, personfit	Item fit indices
LRtest, plotGOF, Waldtest	Model tests
NPtest	Non parametric tests on local stochastic independence, multidimensionality and learning

Name (difR)	Function
difLogistic	Logistic regression DIF method (many others available as well)
plot(object, plot="itemCurve", item = x)	Plot ICC of item x

Test Construction and Test Analysis with R

STEP 9: SET UP NORM SCORES

Learning objectives

- Norm score modeling on manifest and latent level
- Please keep in mind:
  - Traditional norming requires at least 250 per group, continuous norming already works with 100
  - Norming sample has to representative / stratified

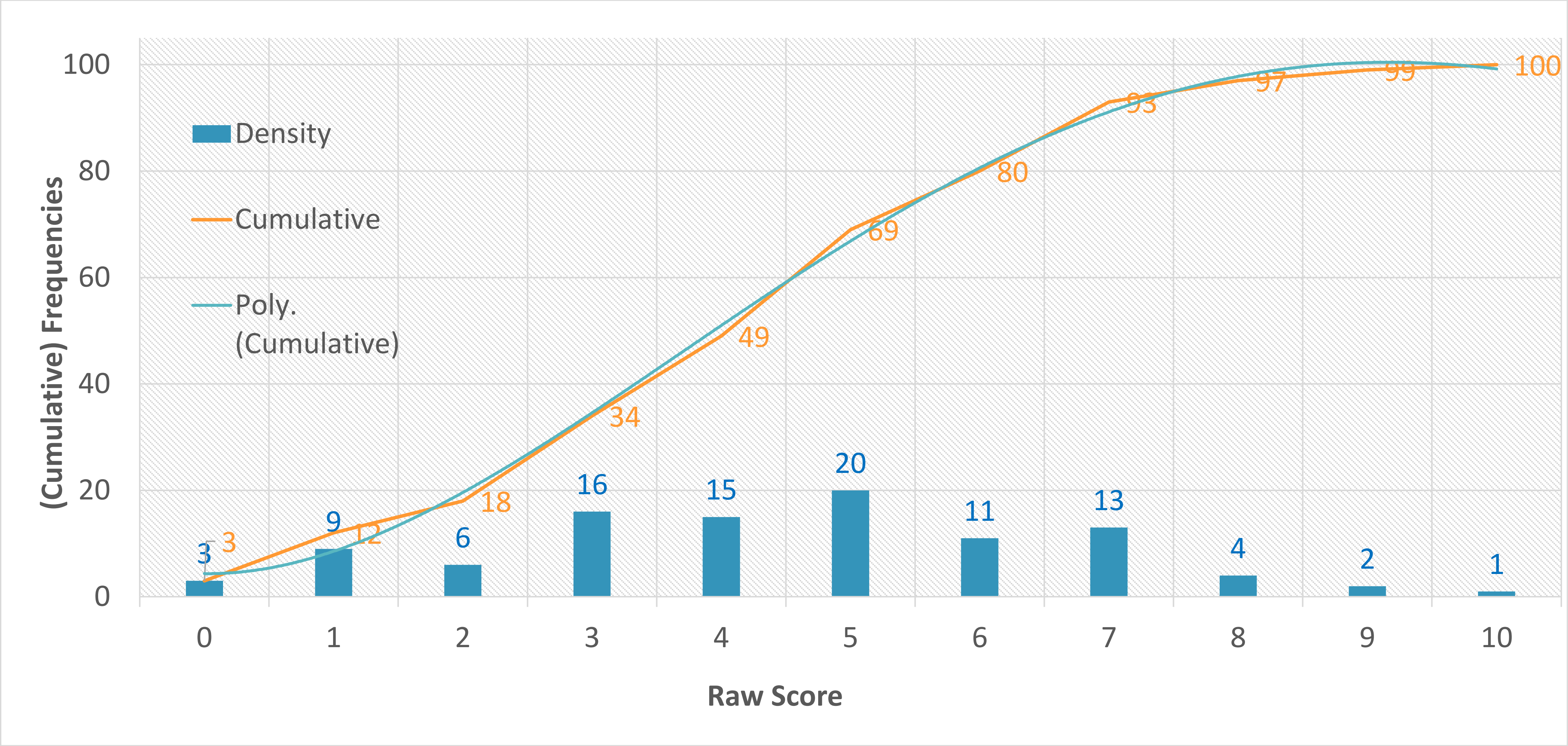
Where to look for help?

- https://www.psychometrica.de/cNorm_en.html
- vignette("cNORM-Demo", package = "cNORM")
- Lenhard, W., & Lenhard, A. (2021). Improvement of Norm Score Quality via Regression-Based Continuous Norming. *Educational and Psychological Measurement, 81*(2), 229–261. https://doi.org/10.1177/0013164420928457

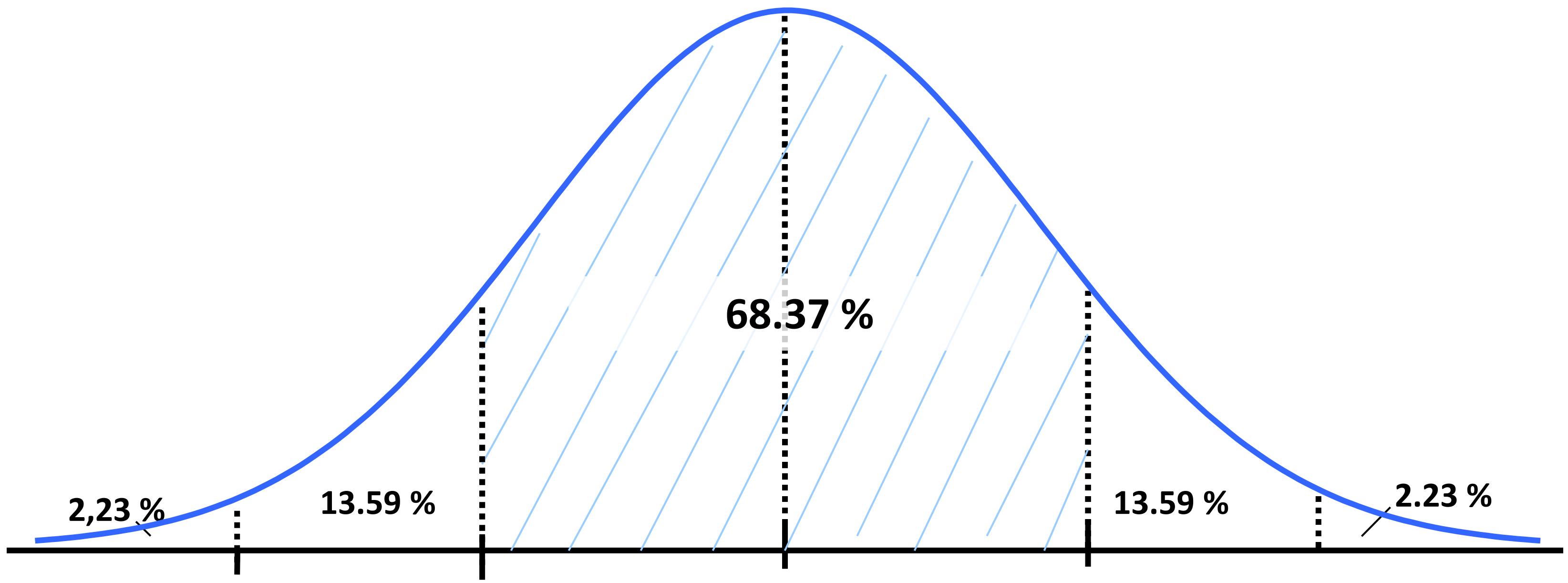
Basic functions you'll need:
cNORM package

Name	Function
dataset ‘elfe’	Representative sample of reading comprehension test results from grade 2 to 5
dataset ‘CDC’	Body weight, height and BMI in boys and girls age 2 to 25 (N = 45 035)
rankByGroup	Determine manifest percentiles and norm scores
cnorm	Set up regression based norming model
plot	Different plots: „norm“, „raw“, „percentiles“, „series“, „subset“, „derivative“ …
normTable, rawTable	Get norm tables

Frequency Distribution: Density and Cumulative Distribution Function (CDF)



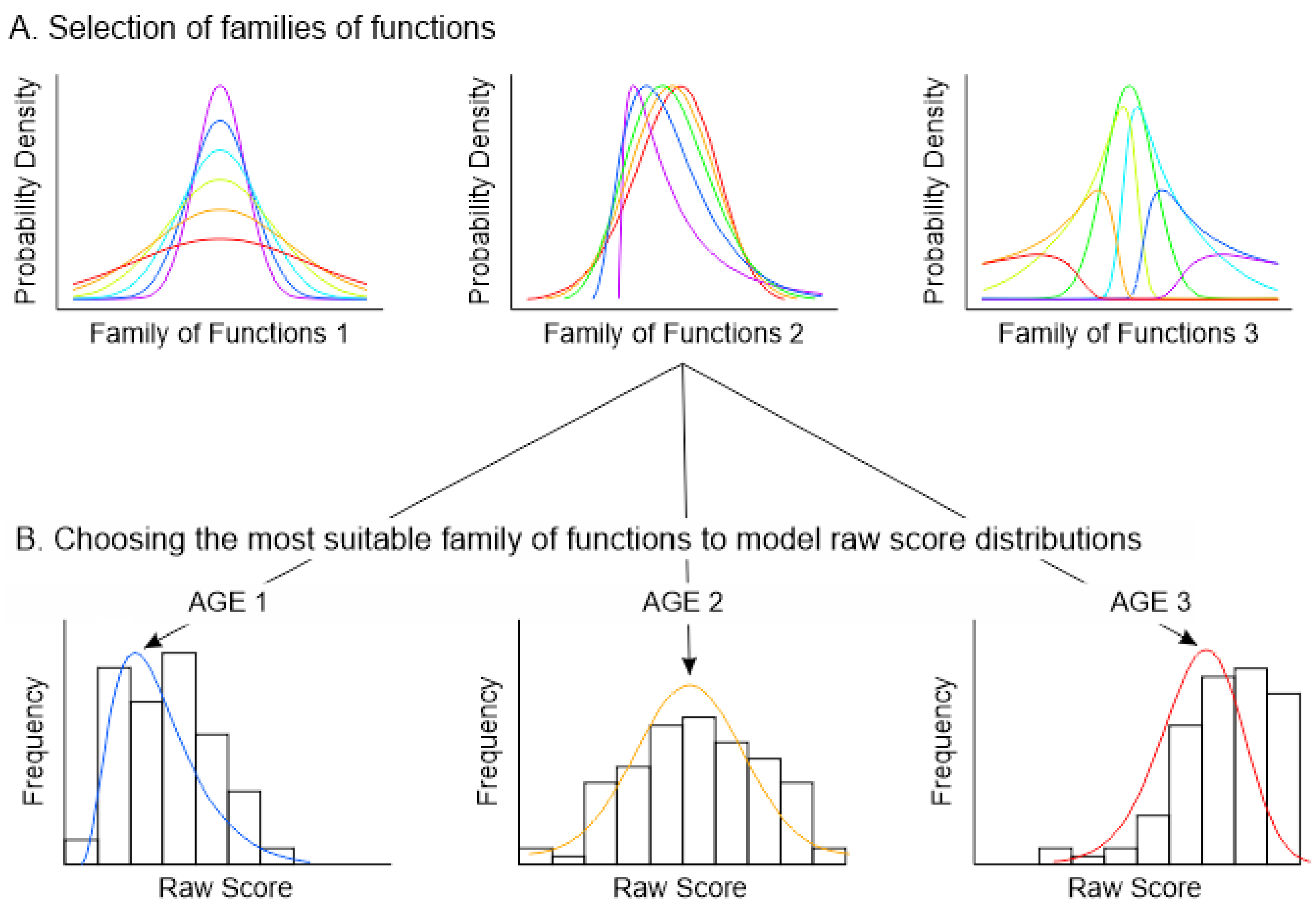
Norm scales



z	- 2	- 1	0	+ 1	+ 2
T	**30**	**40**	**50**	**60**	**70**
IQ	70	85	100	115	130
Percentile (PR)	**2.23 %**	**15.81 %**	**50.0 %**	**84.18 %**	**97.77 %**

Parametric Modeling over Age

Example 3 Parameteric Model of the Box-Cox-Family (LMS sensu Cole & Green, 1990)

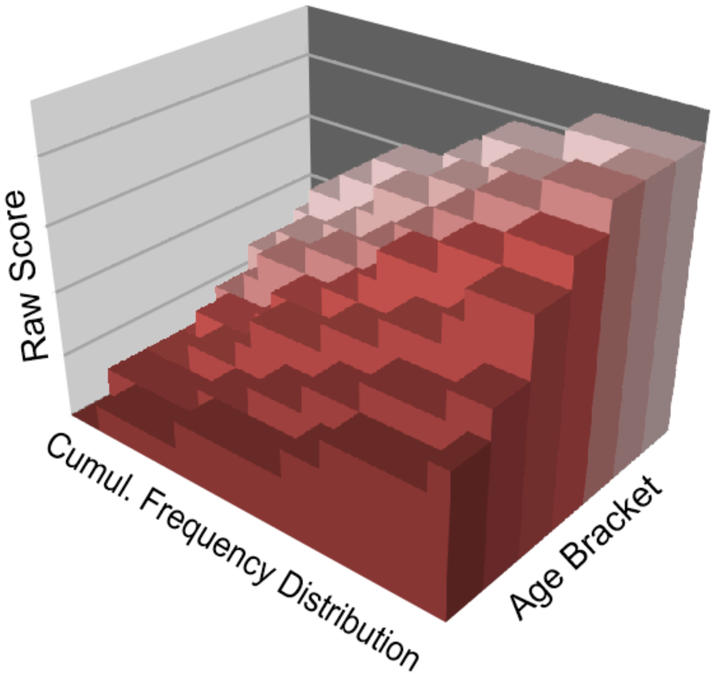


Lenhard, Lenhard & Gary, 2019

Semi-parametric modeling of raw score, age and location via higher order Taylor polynomials

Package cNORM

A. 3D Raw score distribution (observed)



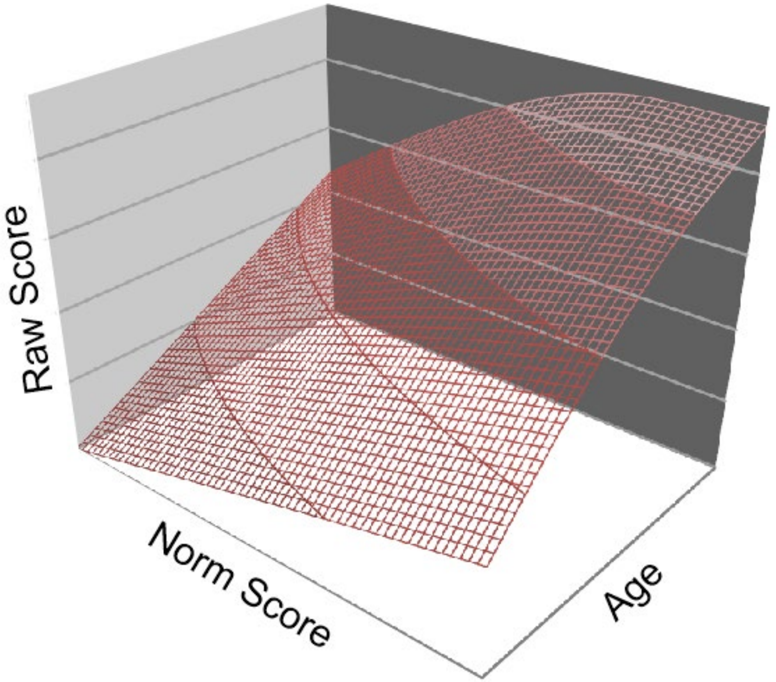
B. Multiple regression

Inter-cept	a	a^2	a^3	a^4	$\cdots$
l	$l\,a$	$l\,a^2$	$l\,a^3$	$l\,a^4$	$\cdots$
l^2	l^2a	l^2a^2	l^2a^3	l^2a^4	$\cdots$
l^3	l^3a	l^3a^2	l^3a^3	l^3a^4	$\cdots$
l^4	l^4a	l^4a^2	l^4a^3	l^4a^4	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

= significant term in the multiple regression

$$r = c_0 + c_1\,l + c_2\,l\,a + c_3\,l^4a^4$$

C. 3D Norming model



Lenhard, Lenhard & Gary, 2019

Test Construction and Test Analysis with R

Step 10: Measurement invariance

Learning objectives

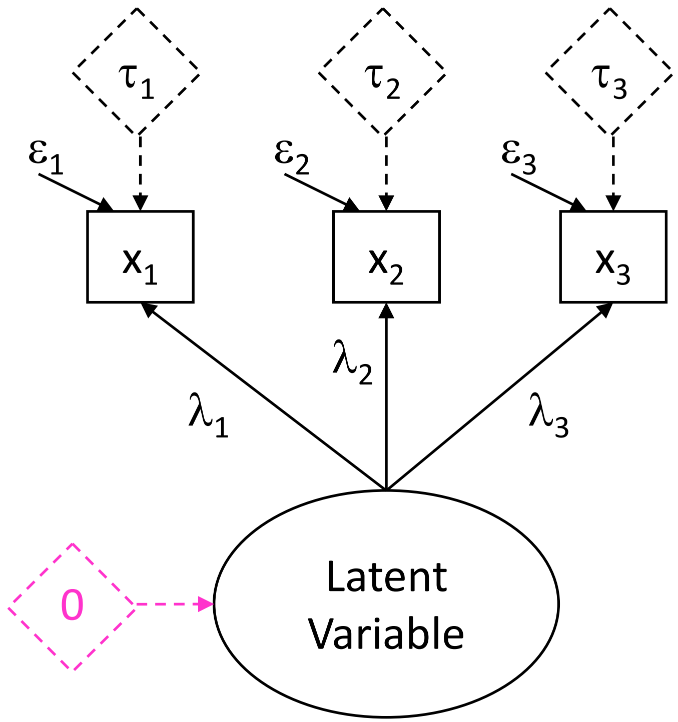
- Basic steps in CFA / SEM and measurement invariance testing

Where to look for help?

- http://lavaan.ugent.be/tutorial/index.html

Methodology: multi-group confirmatory factor analysis (MGCFA)

- Logic of measurement invariance testing (Meredith, 1993; Vandenberg & Lance, 2000):



1. configural Invariance (common factor structure)

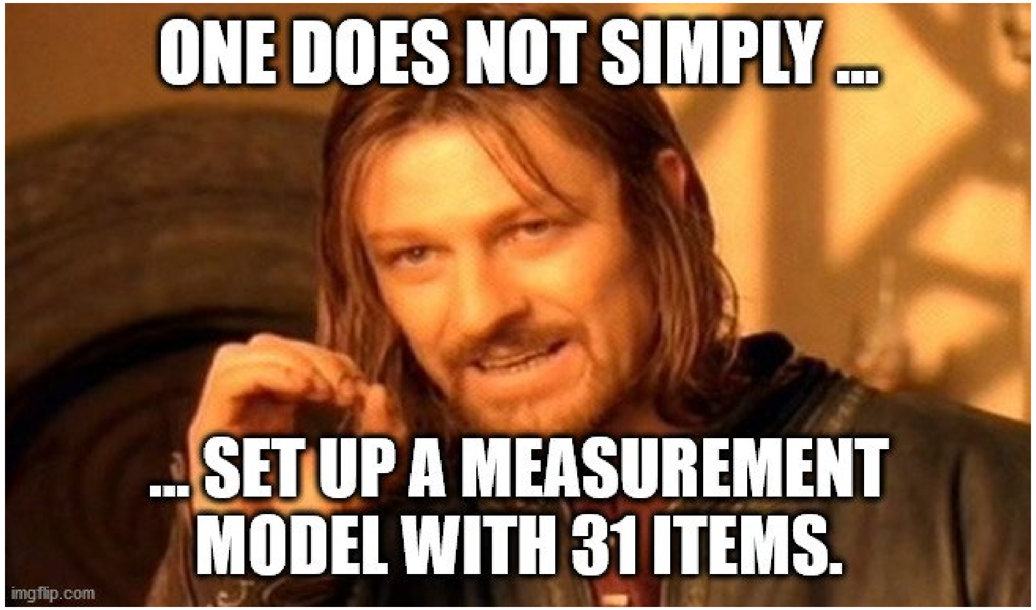
2. Metric Invariance (loadings)

3. Scalar / strong Invariance (loadings + Intercepts)

4. Strict Invariance (Loadings + Intercepts + Residual variances)

Lavaan: Steps

1. Set up measurement model for lavaan
   ⇨ Parcelling (at least 4 parcels)
2. Calculate model (CFA or SEM yield same result), choose estimator WLS for dichotomous items, MLR for continuous
3. Use measurementInvariance-function from semTools package
4. Decision on invariance: $\Delta_{CFI} < .01$ (Cheung & Rensvold, 2002)
   ⇨ Stop when Δ_{CFI} reaches .01



Have a good time and good luck with your work!