Computational Aspects of Psychometric Methods

Ch1: Introduction

NMST570, October 10, 2024

Department of Statistical Modelling
Institute of Computer Science, Czech Academy of Sciences

Institute for Research and Development of Education
Faculty of Education, Charles University, Prague

https://www.cs.cas.cz/martinkova/
@PMartinkova

Table of contents

1. Psychometrics
2. Measurement in social sciences
3. Datasets
4. R
5. Measurement data
6. Conclusion

Psychometrics

Psychometrics is a field of study concerned with the theory and technique of psychological/educational/behavioral measurement.

Field is concerned with objective testing/measurement/assessment of

- skills, knowledge, abilities, educational achievement
- personality traits, attitudes, cognitive function, life satisfaction
- pain, fatigue, depression, anxiety
- physical function, balance, participation,...

What makes behavioral measurement different from measuring temperature, body height, or body weight?

Specifics of Measurement in Social Sciences

What makes behavioral measurement different from measuring temperature or body height:

- It is not possible to directly measure the **construct** of interest (math ability, depression,...), because it is **latent** (unobserved, hidden)
- We use **manifest variables**, e.g. item responses on a questionnaire, to make inferences on these latent variables
- The measurement error is omnipresent and has to enumerated and accounted for
- Rather than single manifest variable, we typically use larger number of variables measuring the same latent construct, using multi-item measurement instruments.

History of Psychometrics

- Carl Friedrich Gauss (1777–1855): measurement errors in the context of astronomy and orbital prediction
- Friedrich Bessel (1784–1846): different observers determine slightly different transition time of stars
- Sir Francis Galton (1822–1911): ”regression towards the mean”, Psychometrics is ”art of imposing measurement and number upon operations of the mind”
- Charles E. Spearman (1863—1945): general intelligence factor (*g* factor), rank correlation coefficient, correction for attenuation

History of Psychometrics (cont.)

- Alfred Binet and Théodore Simon: Binet-Simon IQ test, 1905 concept of mental age
- Lewis M. Terman: IQ score (quotient of mental age divided by chronological age multiplied by 100)
- WWI: Army Alpha and Beta (E.L. Thorndike, A.S. Otis, L.L. Thurstone)
- 1935 Founding of the Psychometric Society
- 1936 First issue of the Psychometrika journal

To learn more, go to `https://www.psychometricsociety.org/history`

Development and validation of measurement instrument

- Defining the construct of interest
- Item writing
- Rater training
- Pretesting
- Administration

- Validation of measurement instrument
- Respondent scaling

Psychometrics involved

- Validation of measurement instrument
  - Gaining proofs of test validity
  - Gaining proofs of test reliability
  - Analyzing item functioning
  - Analyzing impact of group membership and other covariates
- Respondent scaling
- Optimal item selection

The tasks may differ depending on area of measurement...

Educational measurement

- Formative vs. summative, high-stakes vs. low stakes
- Classroom testing
- Language tests (TOEFL, Duolingo English Test)
- State matura examinations
- Standardized admission tests: SAT, MCAT, GMAT, etc.
- National/state testing. US: "No child left behind".
- International large scale assessment: PISA, TIMSS, PIRLS, etc.

For Czech matura data, see `https://www.cermat.cz/`
For more information and data from ILSA, see `https://ilsa-gateway.org/`

Psychological measurement

- Cognitive abilities
  - Binet-Simon IQ test (1905)
  - World War II: the Army General Classification Test
- Personality traits
  - Big 5 (OCEAN): openness, conscientiousness, extroversion, agreeableness, and neuroticism
- Emotional states
- Attitudes
- Diagnosis of psychological disorders

For raw datasets, see `https://openpsychometrics.org/_rawdata/`

Health-related measurement

- SHARE: Survey of Health, Ageing and Retirement in Europe
- PROMIS: Patient-Reported Outcomes Measurement Information System
    - Physical function
    - Pain intensity, pain interference
    - Fatigue
    - Sleep disturbance
    - Depression
    - Anxiety, etc.
- Clinical outcomes in multiple sclerosis (Řasová et al., 2012)
    - Tremor, ataxia, balance, muscle power, etc.

For more information and data from SHARE, see `https://www.share-project.org/`

For more information and data from PROMIS, see `https://www.healthmeasures.net/`

Řasová et al. (2012). Assessment set for evaluation of clinical outcomes in multiple sclerosis: Psychometric properties. *Patient Related Outcome Measures*, 3, 59. doi: 10.2147/prom.s32241

Other behavioral measurement

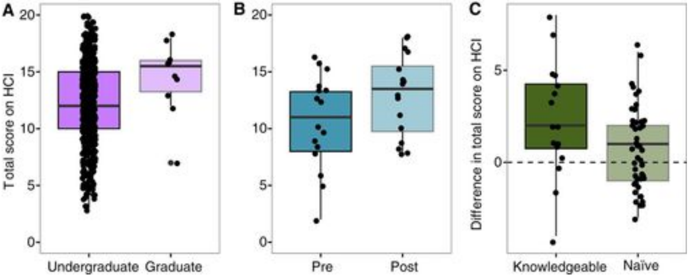
- Grant proposal peer-review (NIH, NHS, GAČR, TAČR)
- Journal peer-review
- Hiring and promotions
- Quality of institutions
- etc.

Selected datasets and sample studies

- Homeostasis Concept Inventory (HCI)
- Grant proposal peer review (AIBS and NIH)
- Graduate Management Admission Test (GMAT)
- Learning to Learn
- Attitudes towards Expulsion of Sudeten Germans
- Medical school admission test

Homeostasis Concept Inventory (HCI)

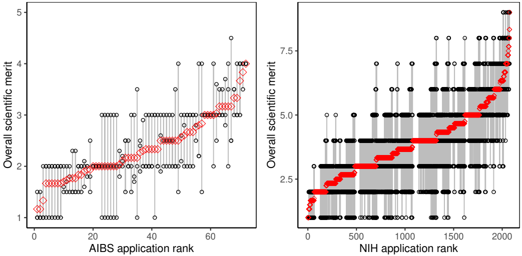
- Conceptual assessments in physiology
- Development and validation of conceptual frameworks, and conceptual assessments
- Validation of HCI



McFarland et al. (2017). Development and Validation of the Homeostasis Concept Inventory. CBE-LSE, 16(2), ar35. doi 10.1187/cbe.16-10-0305

Grant proposal peer review

- American Institutes of Biological Sciences (AIBS) and National Institute of Health (NIH) grant proposal peer-review data
- Originally used to explain zero reliability under restricted samples



Erosheva, Martinková, and Lee (2021). When zero is not zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *JRSS-A*. https://doi.org/10.1111/rssa.12681

Interactively: Reliability section in https://shiny.cs.cas.cz/ShinyItemAnalysis/

Height Inventory

- Item analysis, scaling with multi-item measurements

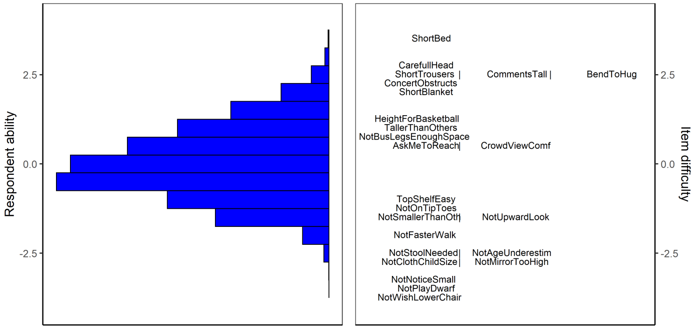
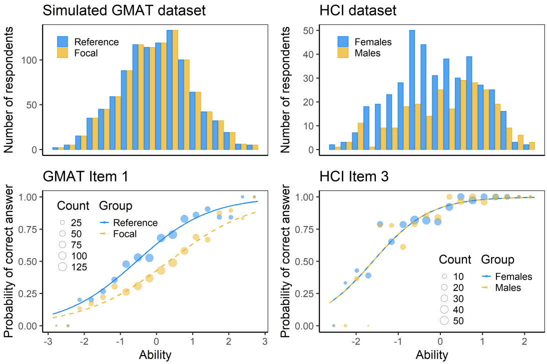


Figure 1: Item-person map for the `HeightInventory` dataset.

Rečka & Cígler. Height Inventory: A psychological attempt at measuring body height. Unpublished manuscript.

Graduate Management Admission Test (GMAT)

- Differential item functioning (DIF) analysis may provide deeper understanding to test functioning among groups.



Simulated GMAT dataset

HCI dataset

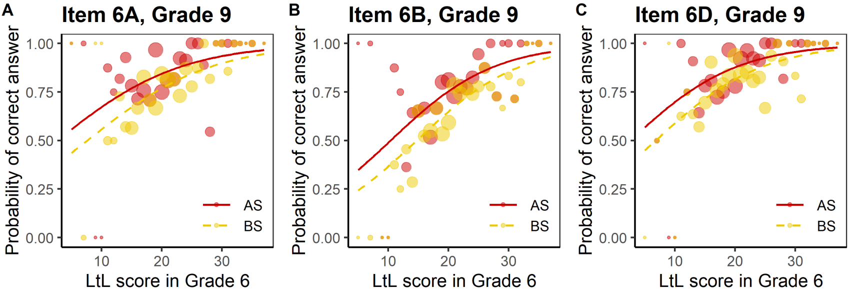
GMAT Item 1

HCI Item 3

Martinková et al. (2017). Checking Equity: Why DIF Analysis should be a Routine Part of Developing Conceptual Assessments. CBE-LSE, 16(2), rm2. doi 10.1187/cbe.16-10-0307

Learning Competence

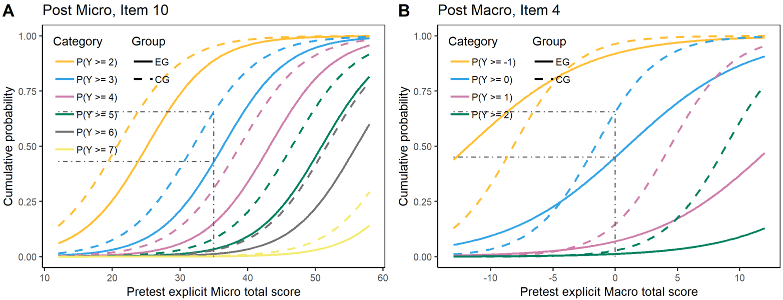
- DIF-C can provide proof of instructional sensitivity, even when differences in change are not visible in total scores.



Martinková, Hladká, and Potužníková (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, 66, 101286. doi: 10.1016/j.learninstruc.2019.101286

Attitudes towards Expulsion of Sudeten Germans

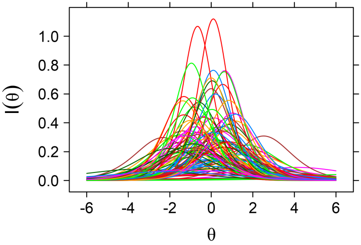
- Change in attitudes towards expulsion after playing a videogame
- DIF-C can provide more detailed treatment sensitivity analysis

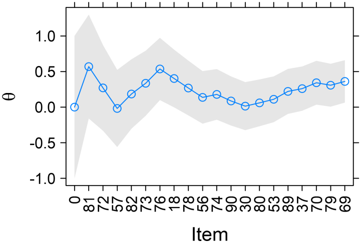


Kolek, Šisler, Martinková, and Brom (2021). Can Video Games Change Attitudes towards History? Results from a Laboratory Experiment Measuring Short- and Long-term Effects. *Journal of Computer Assisted Learning*, 37, 1348–1369. doi: 10.1111/jcal.12575

Medical school admission test

- Simulation studies of new DIF detection methods
- Adaptive selection of optimal items





Drabinová and Martinková (2017). Detection of Differential Item Functioning with Nonlinear Regression: A Non-IRT Approach Accounting for Guessing. *Journal of Educational Measurement*, 54(4), 498–517. doi: 10.1111/jedm.12158

Štěpánek and Martinková (2020). Feasibility of computerized adaptive testing evaluated by Monte-Carlo and post-hoc simulations. *FedCSIS*. doi: 10.15439/2020f197

Software for psychometric analysis

Software for this class:

- R (https://cran.r-project.org/)
  - Psychometric libraries https://CRAN.R-project.org/view=Psychometrics
- RStudio IDE (https://www.rstudio.com/products/rstudio/download/)

Other general statistical software:

- SPSS, Stata (glamm), SAS, JASP, Jamovi, etc.

IRT software

- IRTPRO, flexMIRT, Winsteps, ConQuest, etc.

Software for factor analysis, structural models:

- Mplus, Lisrel, AMOS, etc.

Software for traditional item analysis

- Iteman, Lertap, etc.

ShinyItemAnalysis

ShinyItemAnalysis is an R package and an online application for

- teaching/learning psychometric methods (CTT, IRT, DIF,...)
- complex analysis of educational and psychological tests

With the aim of wide-spreading the methodology into distant fields and geographic regions.

See also

- Martinková and Hladká (2023)
- Martinková and Drabinová (2018)

Types of data

- Nominal
  - Unordered categorical variables, used to "name" a series of values/categories without predefined ordering (e.g., type of school, respondent's gender)
- Ordinal
  - Include ordering but provide no information on the distance between subsequent categories (e.g., items on a rating scale: rarely – sometimes – often – always)
- Interval
  - Numerical variables which assume equal differences between successive categories, but have arbitrary position of zero (e.g., temperature)
- Metric
  - Numerical variables which have equal intervals between successive categories and a natural (meaningful) zero (e.g., age, height, weight)

Types of data in R

Basic data types:

- `numeric`
- `integer`
- `character`
- `logical`

More complex R objects:

- `vector`
- `factors`
- `matrix`
- `array`
- `data.frame`
- `list`

Describing data

- Proportions and ratios
- Mean, modus, median, quartiles, percentiles
- Variance, standard deviation, range, interquartile range

Multiple-choice items

Typical in educational measurement. Example:

$$-6-(-10)=?$$

a. -16

b. -4

c. 4

When the key is provided, this type of items can be scored as

a. FALSE (0)

b. FALSE (0)

c. TRUE (1)

Ordinal items

Partial-credit items in educational measurement. Example:

"In what way are an orange and a banana alike?"

2pts: Provides pertinent general categorization (e.g., "Both are fruit.")

1pts: Provides one or more common properties (e.g., "Both are food.")

0pts: Provides specific properties for each member of the pair, or a wrong answer (e.g., "Both are round.")

Typical in psychological and health-related assessment. Items rated on a rating scale, such as:

1pt: Rarely

2pts: Sometimes

3pts: Often

4pts: Always

Continuous items

- Example: the Eysenck personality inventory
  - "Do you often long for excitement?", etc.
  - Checks on a 112 mm segment line, with 0 corresponding to "Almost never" and 112 corresponding to "Almost always".
- Item response time

Test scores

- Total scores (raw scores)
- Success rate (percentage of correct answers)
- Z-scores
- T-scores
- Percentiles

Binary variables

Defined by probability mass function $\mathrm{P}(X = x_i)$

Mean and standard deviation:

$$\mathrm{E}(X) = \sum_i x_i \mathrm{P}(X = x_i) \qquad \operatorname{var}(X) = \sum_i (x_i - \mathrm{E}(X))^2 \mathrm{P}(X = x_i)$$

- Bernoulli distribution

$$\mathrm{P}(X = 1) = \pi$$
$$\mathrm{P}(X = 0) = 1 - \pi$$

- Binomial distribution

$$\mathrm{P}(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}.$$

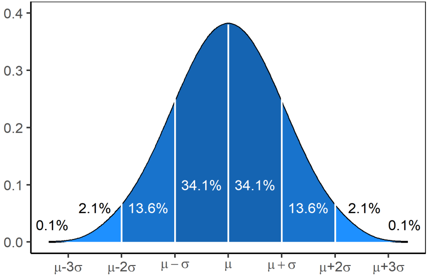
Continuous variables

Defined by probability density function $f_X(t)$ so that

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t. \tag{1}$$

- Normal distribution

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\,\mathrm{d}t.$$



Vocabulary

- Psychometrics
- Construct
- Latent variable
- Manifest variable
- Multi-item measurement
- Nominal – ordinal – interval – metric
- Mean – median – modus
- Variance – standard deviation – (interquartile) range
- Raw scores – Z-scores – T-scores – percentiles
- Distribution
    - discrete: Bernoulli – binomial – uniform
    - continuous: Normal – χ^2 – t – F

Tasks and Exercises

- Read Chapter 1 and Appendices A, B, C, and D
- Run the accompanying R code
  - `A-InstallPackages.R`
  - `A-IntroductionToR.R`
  - `Ch1-Introduction.R`
- Complete Exercises in Appendix E.1
- Explore CZmatura data with `ShinyItemAnalysis` and R
- Search datasets for the project
  - Next time: Share with the class, which data you would like to analyze for your project.
    - What are the items in the measurement instrument?
    - How would you assess reliability and validity?
    - Which groups of respondents would be of interest for between-group comparison?
    - What other questions may be of interest?

Thank you for your attention!

www.cs.cas.cz/martinkova

martinkova@cs.cas.cz

References

Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement*, *54*(4), 498–517. doi: 10.1111/jedm.12158

Erosheva, E. A., Martinková, P., & Lee, C. J. (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *00*, 1—16. doi: 10.1111/rssa.12681

Kolek, L., Šisler, V., Martinková, P., & Brom, C. (2021). Can video games change attitudes towards history? results from a laboratory experiment measuring short- and long-term effects. *Journal of Computer Assisted Learning*, *37*(5), 1348–1369. doi: 10.1111/jcal.12575

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503–515. doi: 10.32614/rj-2018-074

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE-Life Sciences Education*, *16*(2), rm2. doi: 10.1187/cbe.16-10-0307

Martinková, P., Hladká, A., & Potužníková, E. (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, *66*, 101286.

Martinková, P., & Hladká, A. (2023). *Computational aspects of psychometric methods: With R*. CRC Press. doi: 10.1201/9781003054313

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., ... Wright, A. (2017). Development and validation of the homeostasis concept inventory. *CBE-Life Sciences Education*, *16*(2), ar35. doi: 10.1187/cbe.16-10-0305

Řasová, K., Martinková, P., Vyskotová, J., & Šedová, M. (2012). Assessment set for evaluation of clinical outcomes in multiple sclerosis: Psychometric properties. *Patient Related Outcome Measures*, *3*.

Štěpánek, L., & Martinková, P. (2020). Feasibility of computerized adaptive testing evaluated by monte-carlo and post-hoc simulations. In *Proceedings of the 2020 federated conference on computer science and information systems (FedCSIS)* (pp. 359–367). doi: 10.15439/2020f197