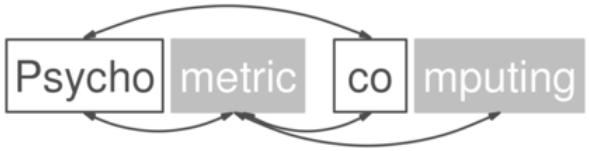
University of Zurich[UZH]

Department of Psychology - Psychological Methods, Evaluation and Statistics

Score-Based Measurement Invariance Tests for Multistage Testing (A Tale of Two and a Half Tests)

Rudolf Debelak, Dries Debeer



Road Map

- What are score-based DIF tests?
- Adaptive Testing: MSTs (and CATs)
- Two and a half solutions
- A simulation study
- Summary and future work

What are score-based tests for DIF?

Score-based DIF tests detect an instability of item parameters with regard to a person covariate:

- Age
- Native language
- Gender
- …

What are score-based tests for DIF?

- **Bradley-Terry Models** (Strobl, Wickelmaier & Zeileis, 2011).
- **Factor analytical models** (Merkle & Zeileis, 2013; Merkle, Fan & Zeileis, 2014)
- **Rasch models** (Strobl, Kopf & Zeileis, 2015; Komboz, Strobl & Zeileis, 2016)
- **Normal-ogive IRT models** (Wang, Strobl, Zeileis & Merkle, 2017)
- L**ogistic IRT models**(Debelak & Strobl, 2018)

What are score-based tests for DIF?

Consider a statistic of model bias B_i on the person level for each item parameter. We assume that under the null model:

- Its expected value for any person $E(B_i)$ is 0.
- This statistic is independent and identically distributed for all test takers.

We now consider sums $\sum B_i$ over sufficiently large groups of test takers.

What are score-based tests for DIF?

Consider a statistic of model bias B_i on the person level for each item parameter. We assume that under the null model:

- Its expected value for any person $E(B_i)$ is 0.
- This statistic is independent and identically distributed for all respondents.

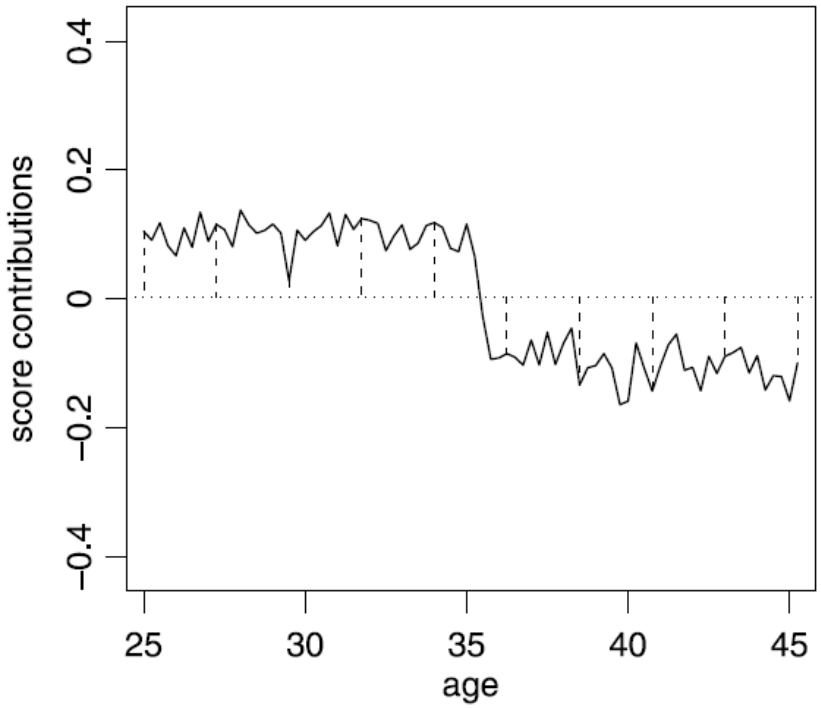
We now consider sums $\sum B_i$ over sufficiently large groups of test takers.

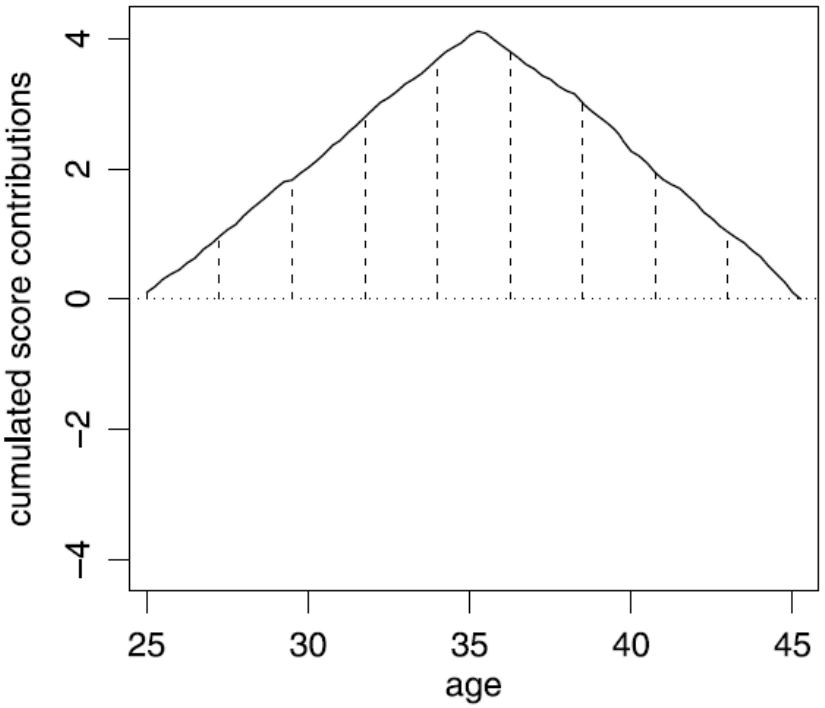
If our null model is correct,

- $\sum B_i$ follows a normal distribution (Central Limit Theorem)
- The related stochastic process is a Brownian bridge (Functional Central Limit Theorem)

These assumptions are met by individual score contributions for ML estimators (Hjort & Koning, 2002; Zeileis & Hornik, 2007).

What are score-based tests for DIF?





What are score-based tests for DIF?

Summary:

- Obtain ML estimates for the item parameters.
- Calculate the individual score contributions
- Order the persons with regards to a person covariate of interest (gender, age).
- Calculate the cumulative sums with regard to this order.
- Compare the stochastic processes (the scores) with the process assumed under the null models (by some test statistic) for an item of interest

«Can you apply this to adaptive tests in R?»

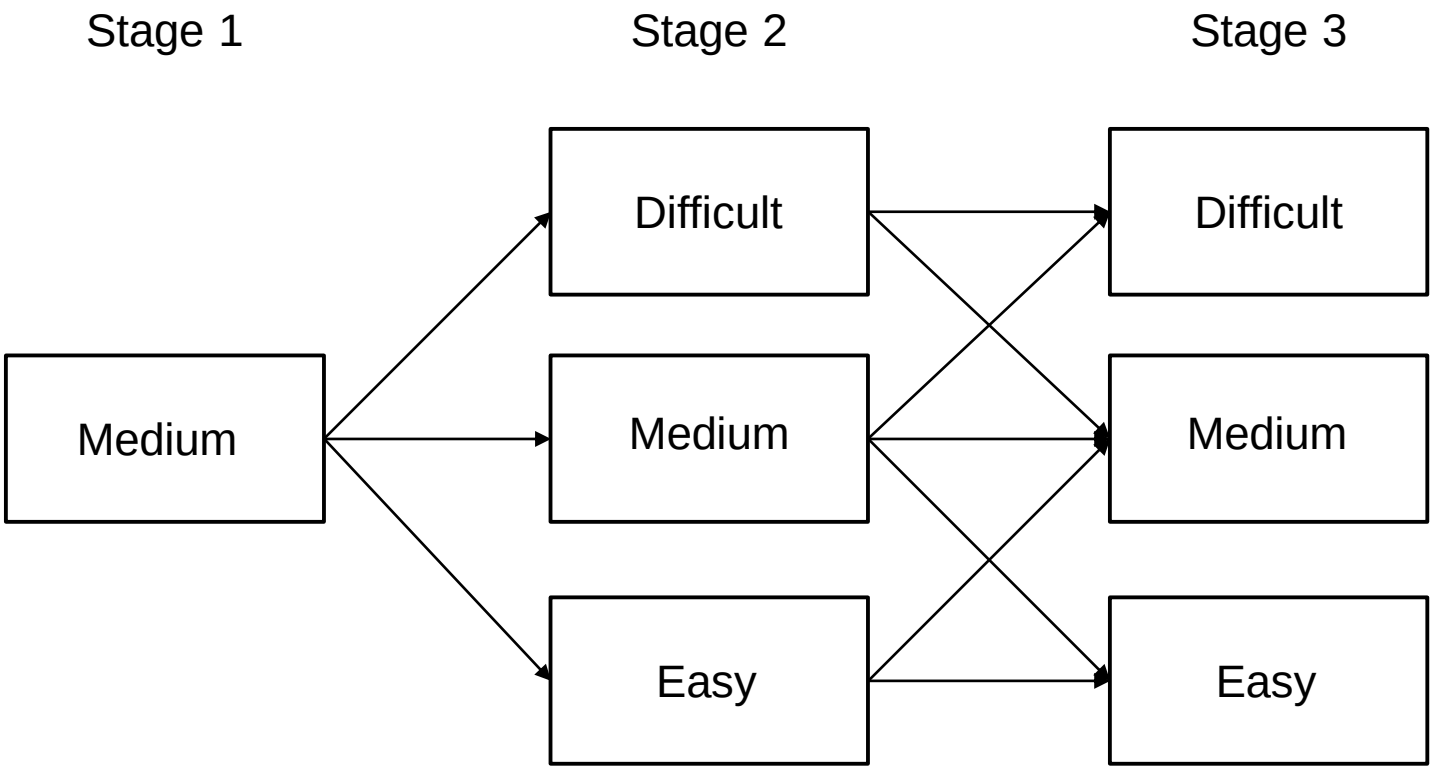
Adaptive Testing: MSTs (and CATs)

- Consider the 2PL model:

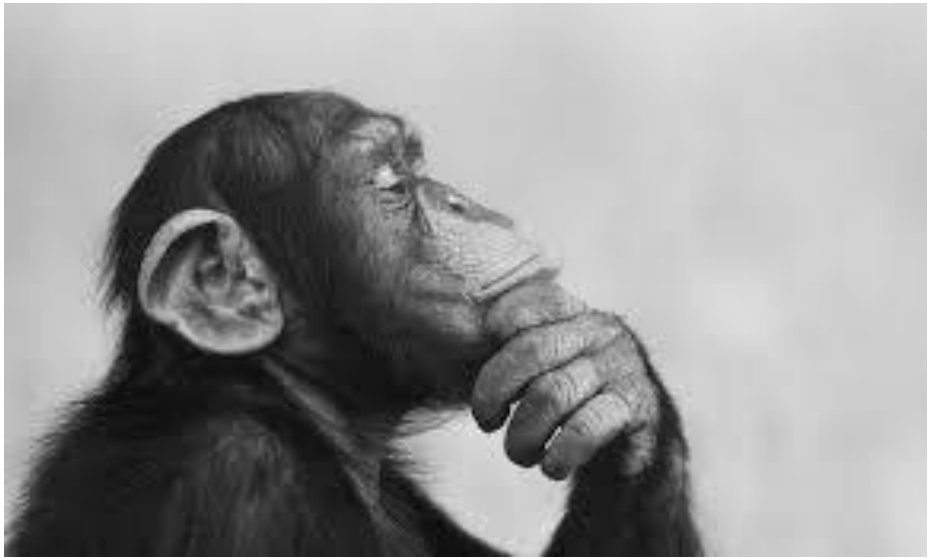
$$P(X_{ij} = 1|\theta_i, a_j, b_j) = \frac{\exp(a_j\theta_i + b_j)}{1+\exp(a_j\theta_i + b_j)}$$

- Further assume that we have a large set of items with known item parameters.

Adaptive Testing: MSTs (and CATs)



«Can you apply this to adaptive tests in R?»



Test 1: Asymptotic Score-Based Tests

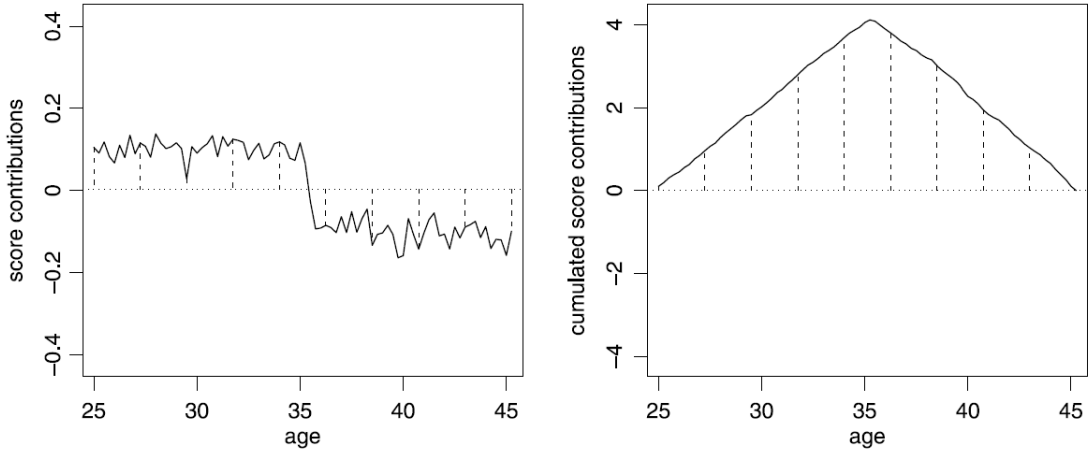
3 Steps:

1. Use the observed data from an adaptive test.
2. Treat the missing data as missing at random and estimate the item parameters.
3. Apply score-based DIF tests for this IRT model.

Test 2: Bootstrap Score-Based Tests

5 Steps:

1. Consider the calibrated item parameters and person parameter estimates
2. For an item of interest, generate artificial responses based on your IRT model and the estimated person parameters.
3. Repeat Step 2 many (e.g., 1000) times.
4. Calculate a score-based statistic of model fit for the original and the artificial data.
5. Calculate p-values.



Bootstrap Score-Based Tests

- ✓ Use calibrated item parameters
- ✓ Use person parameter estimates
- ✓ Calculate p-values based on Bootstrapping (or permutation)

Asymptotic Score-Based Tests

- ✓ Estimate item parameters using an assumed distribution of person parameters
- ✓ Calculate p-values based on asymptotic results.

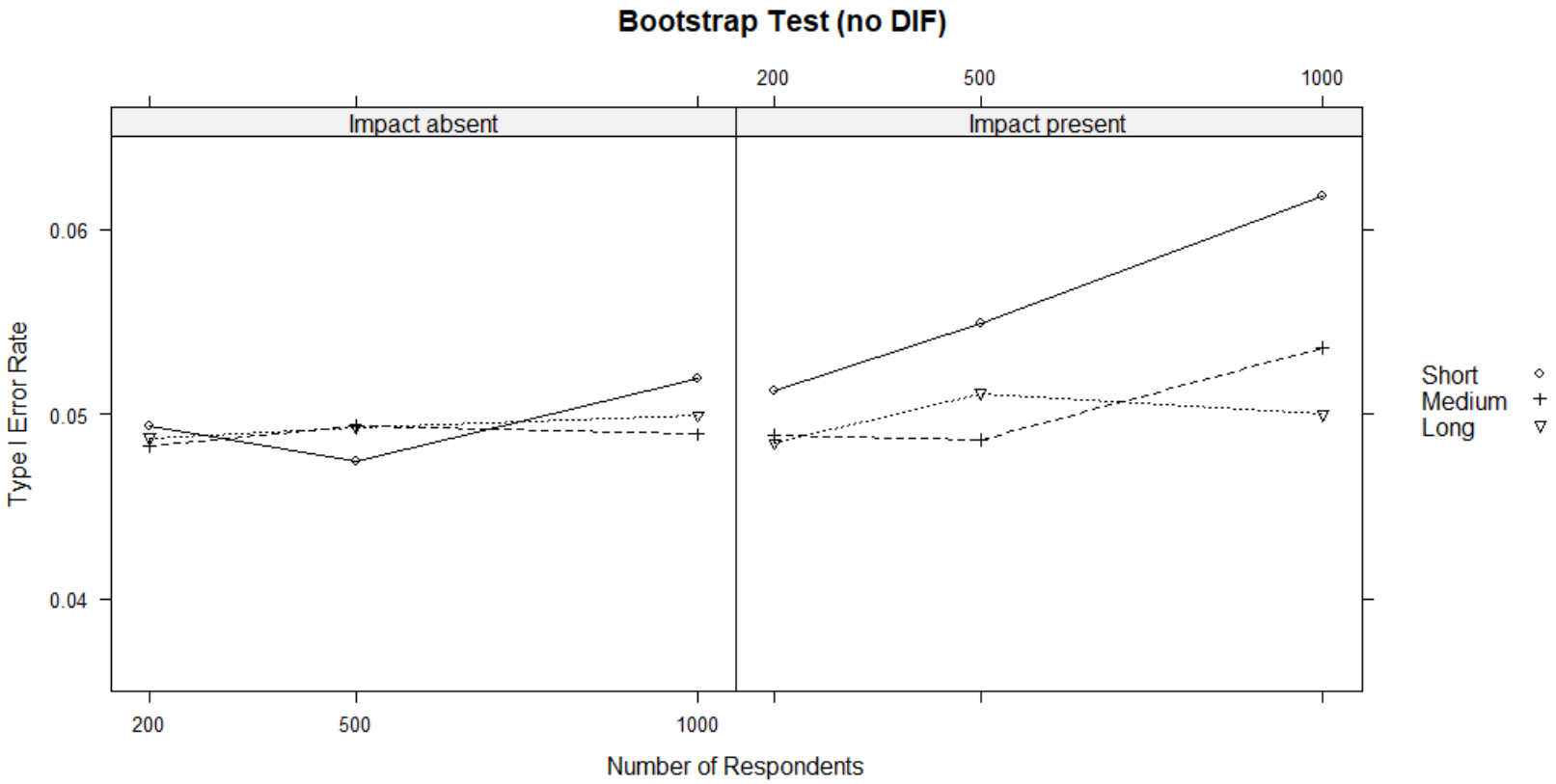
An Evaluation with a Simulation Study

Design:

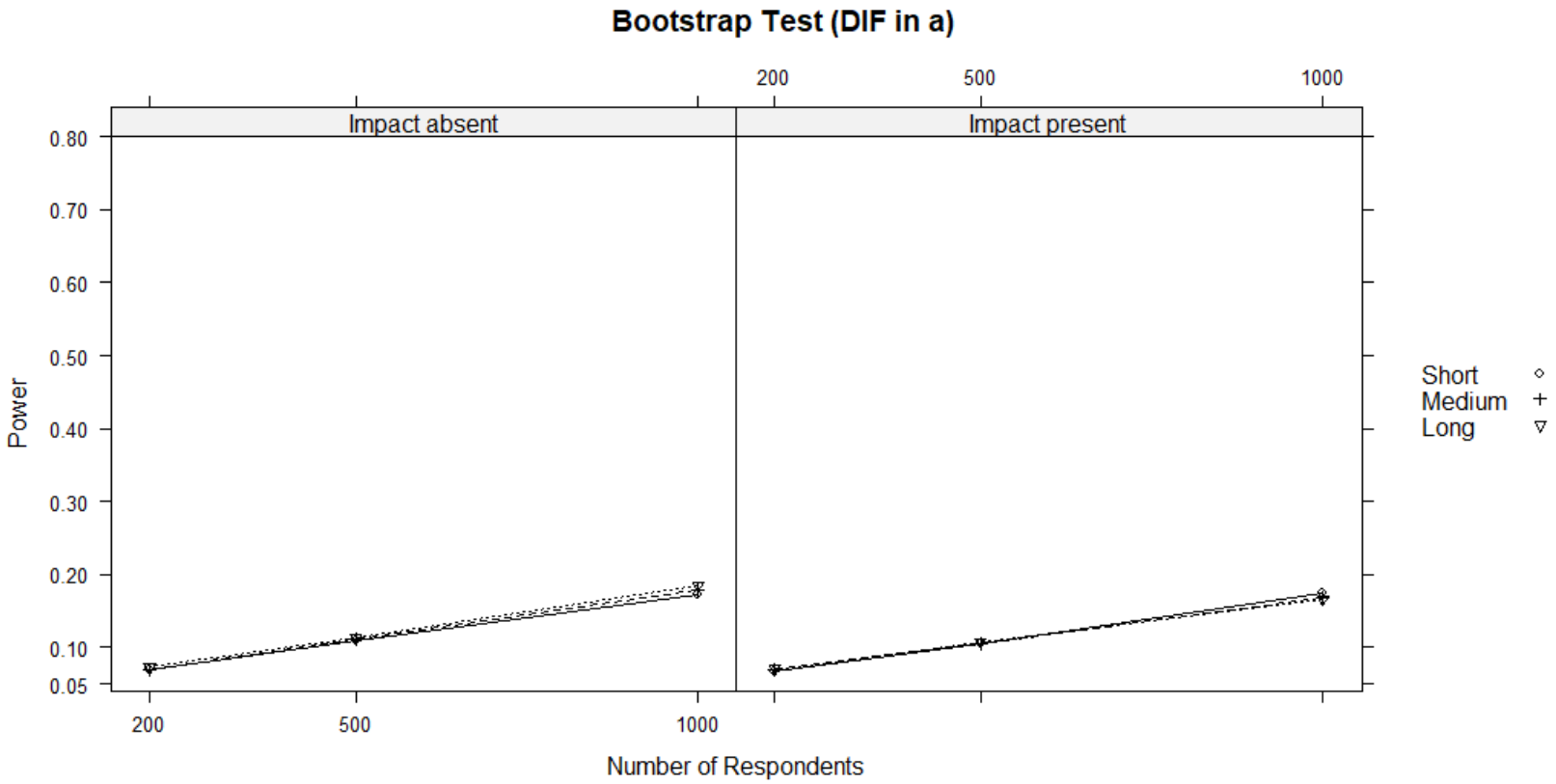
- 1 – 3 – 3 MST design
- 3 sample sizes: 200, 500, 1000 test takers
- 3 lengths of modules: 9, 18, 36 items
- 2PL model
- Two known groups of equal size:
  - Impact absent / present
  - No DIF, DIF of 0.3 in a parameter, DIF of 0.6 in b parameter (4 in 9 items per module)
- Evaluation with Bootstrap score-based tests and asymptotic score-based tests.
- 500 repetitions per condition

Results for the Bootstrap Test

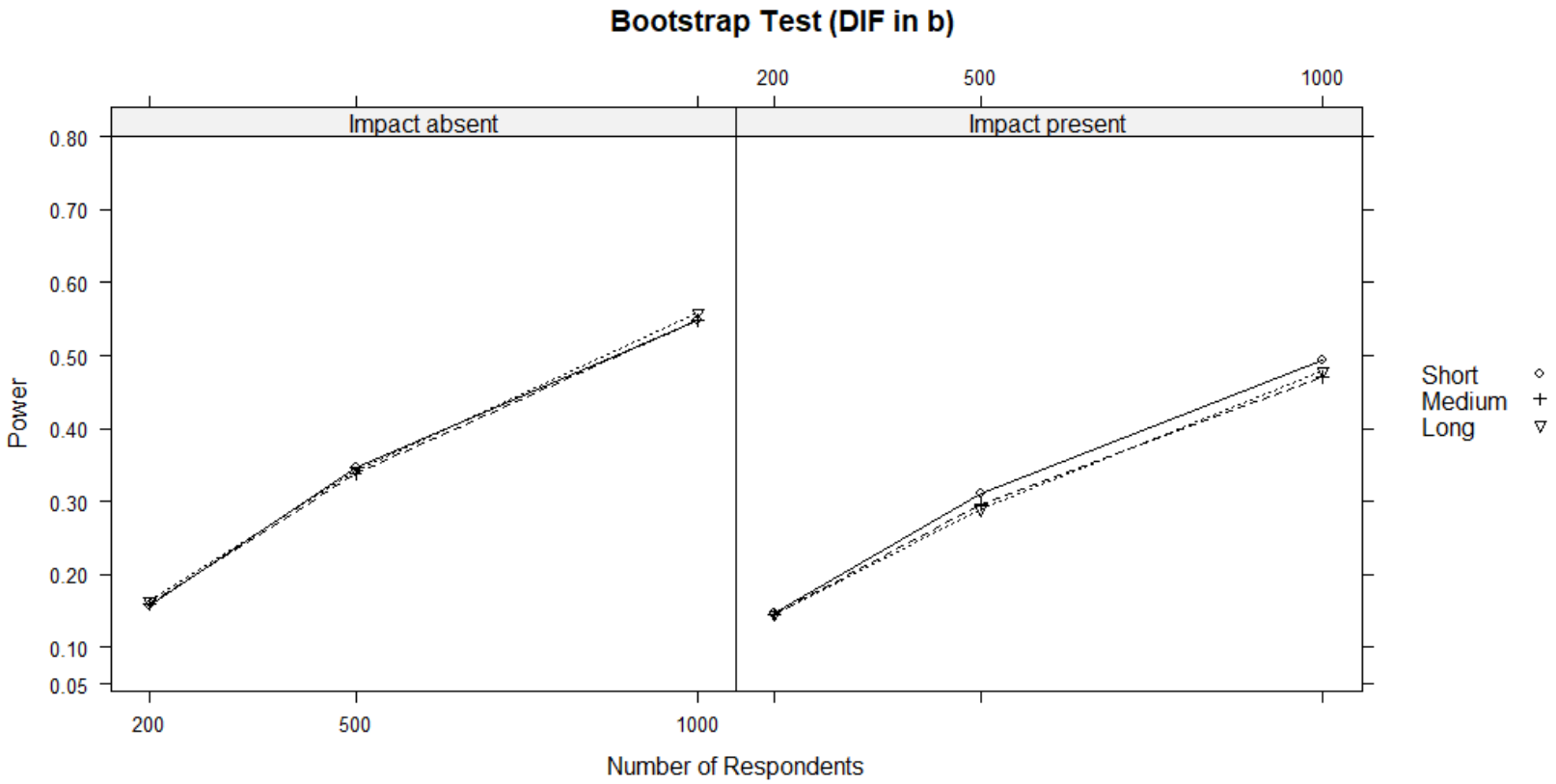
Results for the Bootstrap Test



Results for the Bootstrap Test

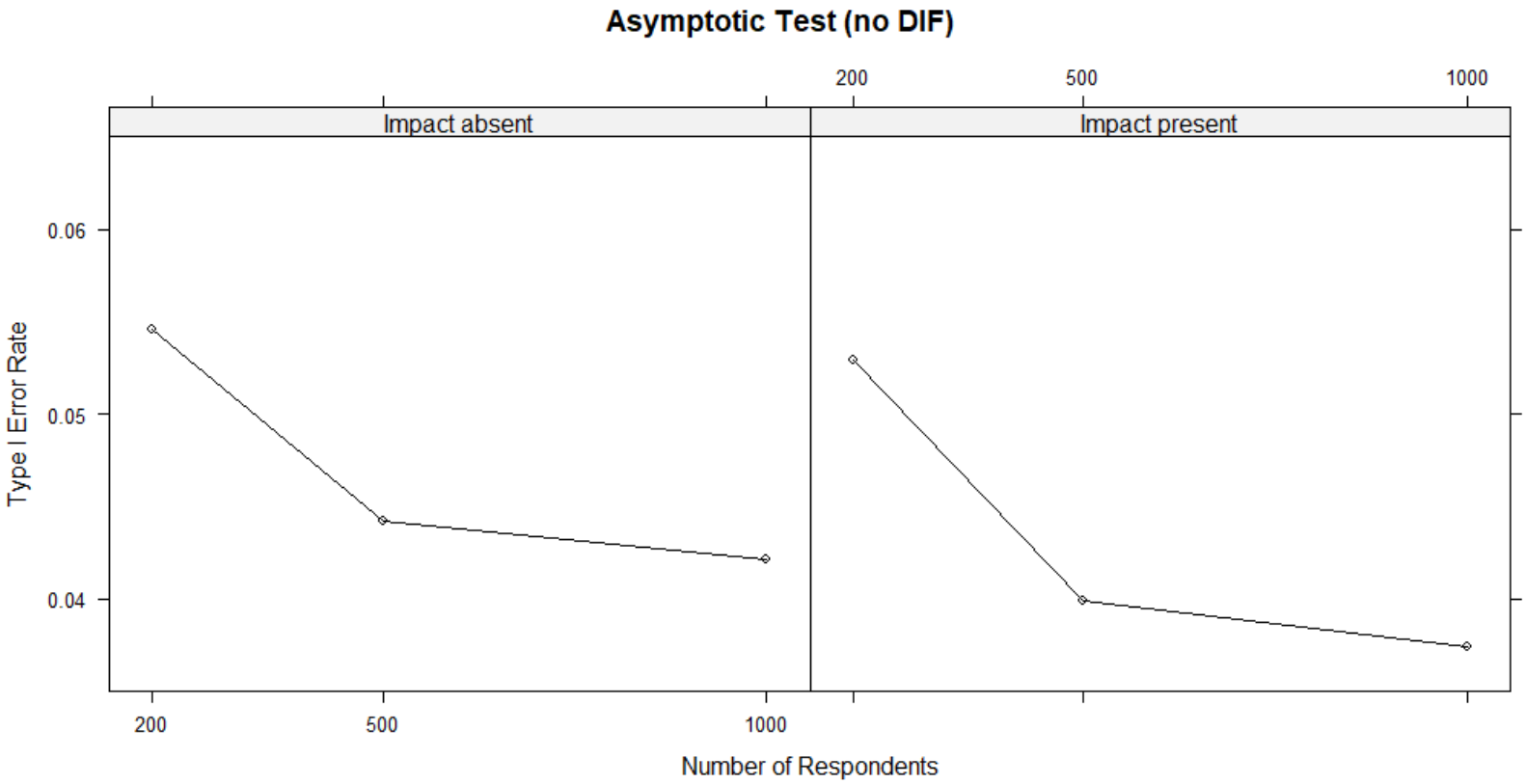


Results for the Bootstrap Test

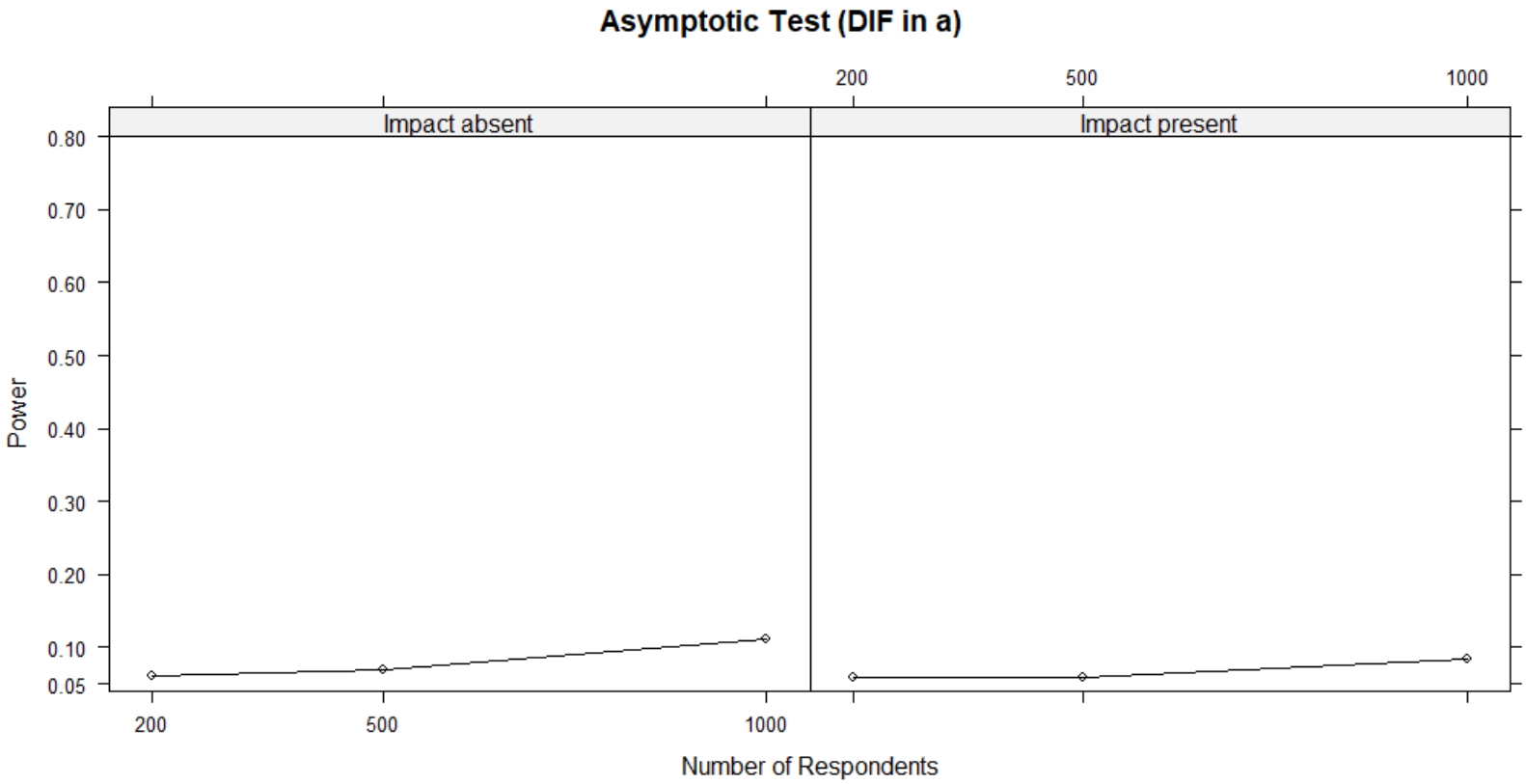


Results for the Asymptotic Test (only short modules)

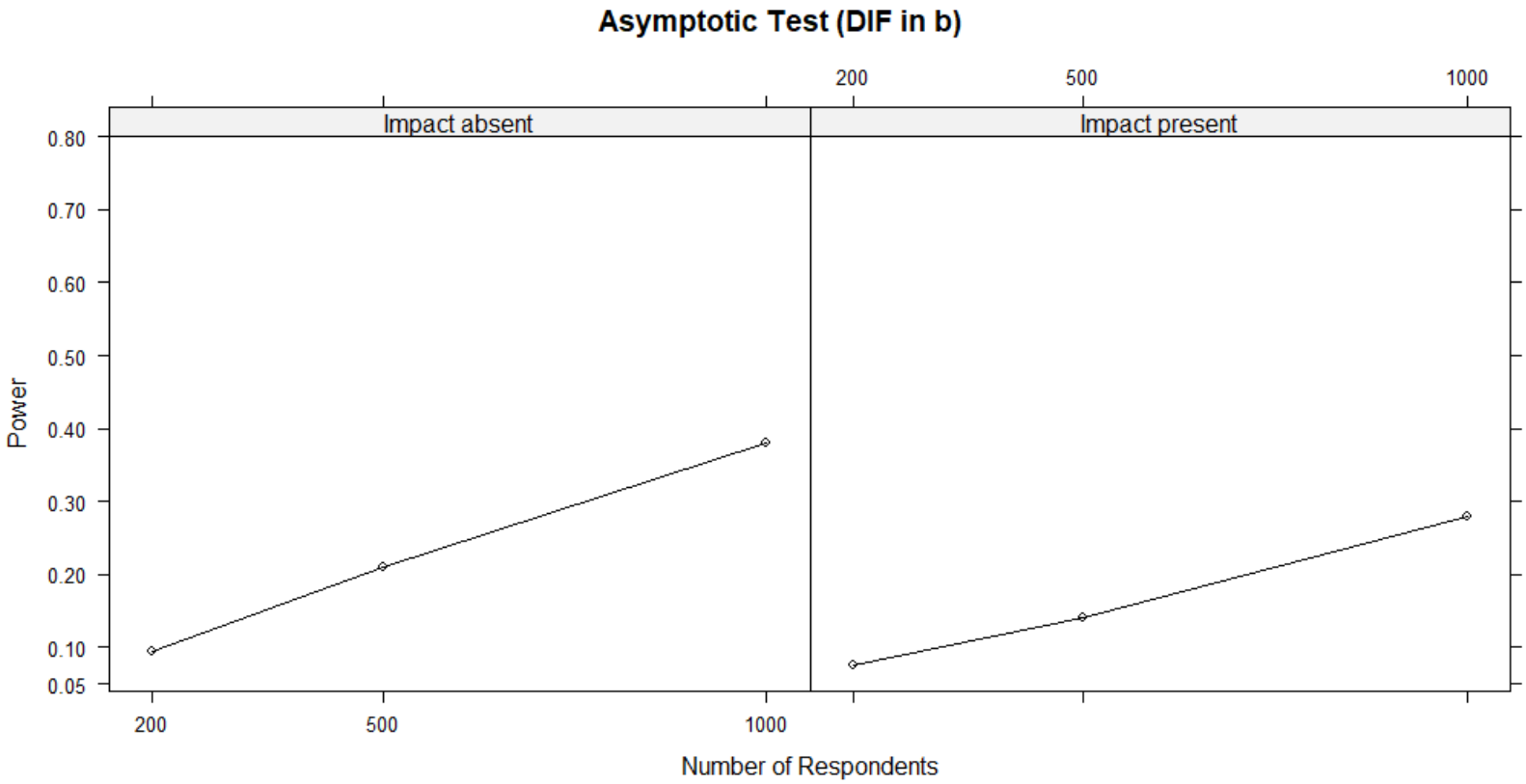
Results for the Asymptotic Test



Results for the Asymptotic Test



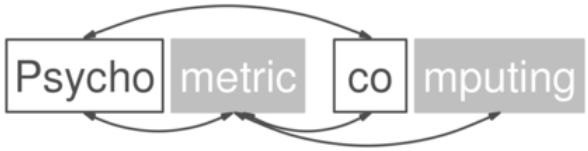
Results for the Asymptotic Test



Summary

- We presented two and a half tests for the flexible detection of DIF in adaptive tests.
- The **Bootstrap score-based test** uses the calibrated item parameters and has **higher power** if these are correct. If not, it shows an increased Type I error.
- The **asymptotic score-based test** estimates the item parameters from the data, which makes it **computationally intensive**.
- A **third approach based on permutation** leads to identical results as the Bootstrap test.
- These and other tests are available in the **mstDIF package** (Debelak, Debeer, & Appelbaum, 2020).

Thank you for your interest!



References

Debelak, R., & Strobl, C. (2018). Investigating Measurement Invariance by Means of Parameter Instability Tests for 2PL and 3PL Models. *Educational and Psychological Measurement*, doi: 10.1177/0013164418777784

Hjort, N. L., & Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, *14*(1-2), 113-132.

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79* (4), 569-584.

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: a generalization of classical methods. *Psychometrika*, *78* (1), 59-82.

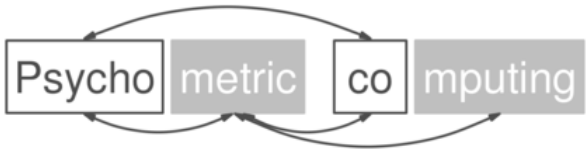
Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80* (2), 289-316.

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics*, *36* (2), 135-153.

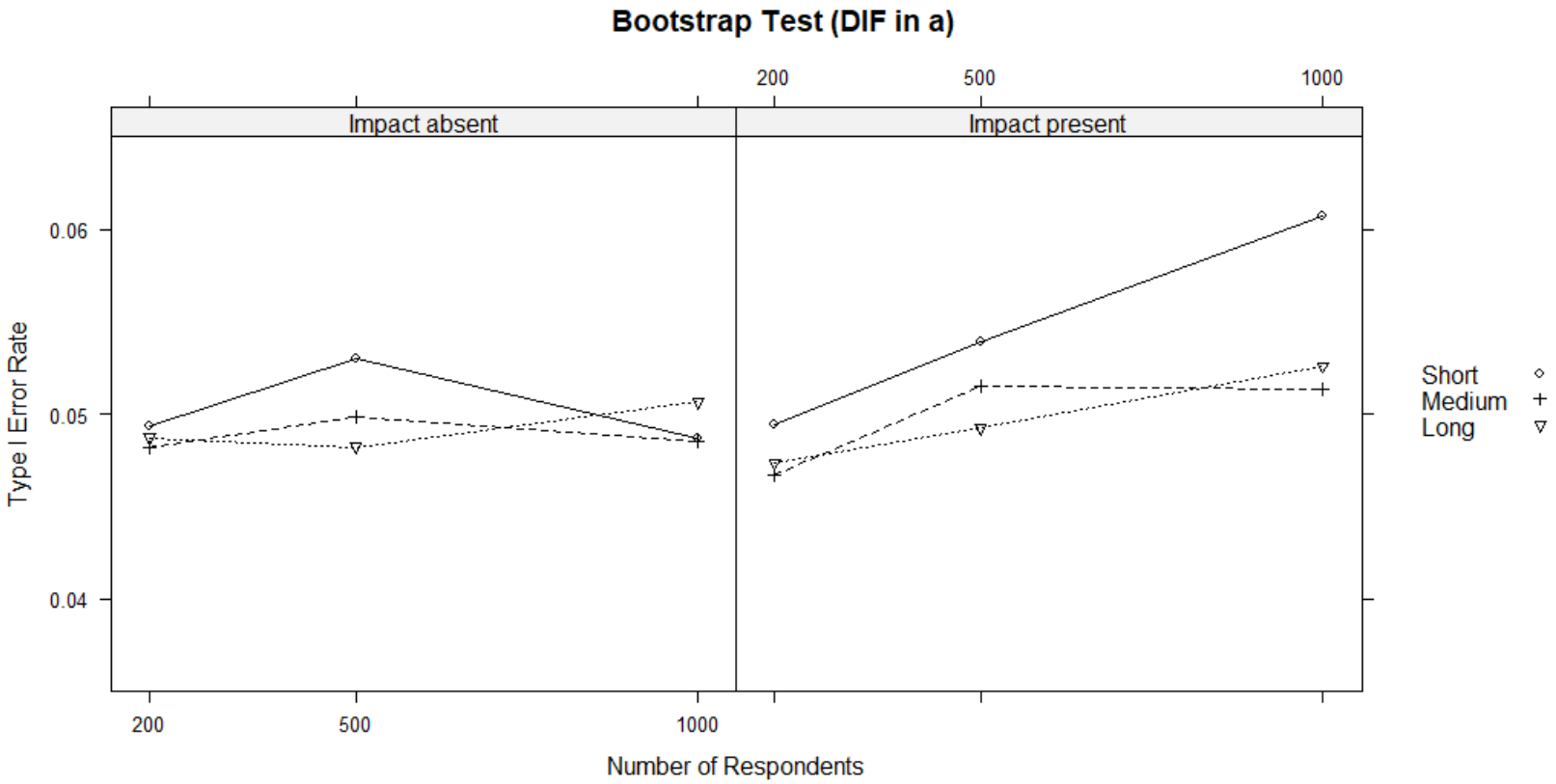
Wang, T., Strobl, C., Zeileis, A., & Merkle, E. C. (2017). Score-based tests of differential item functioning via pairwise maximum likelihood estimation. *Psychometrika*. doi: 10.1007/s11336-017-9591-8

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61* (4), 488-508.

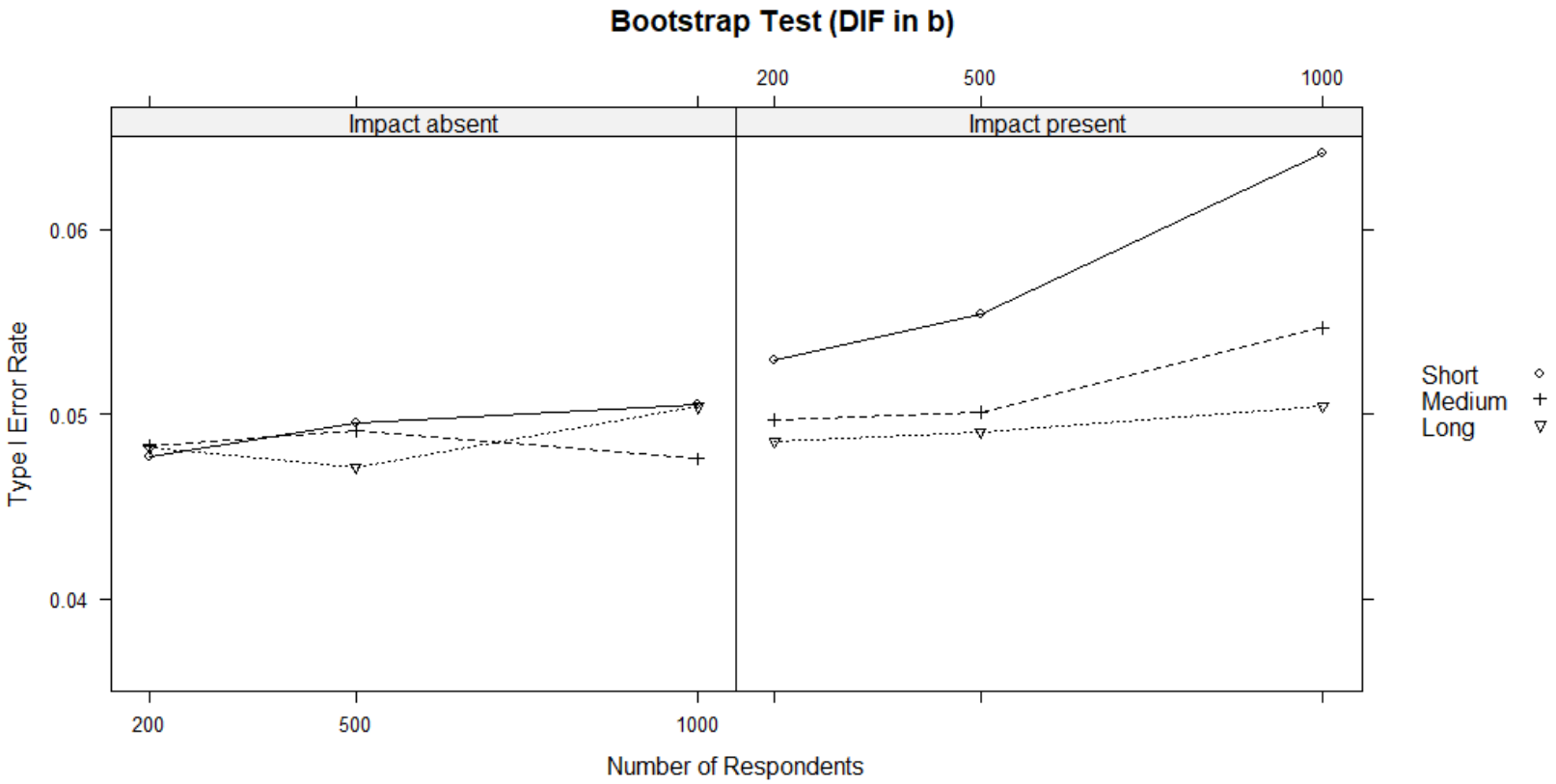
Appendix



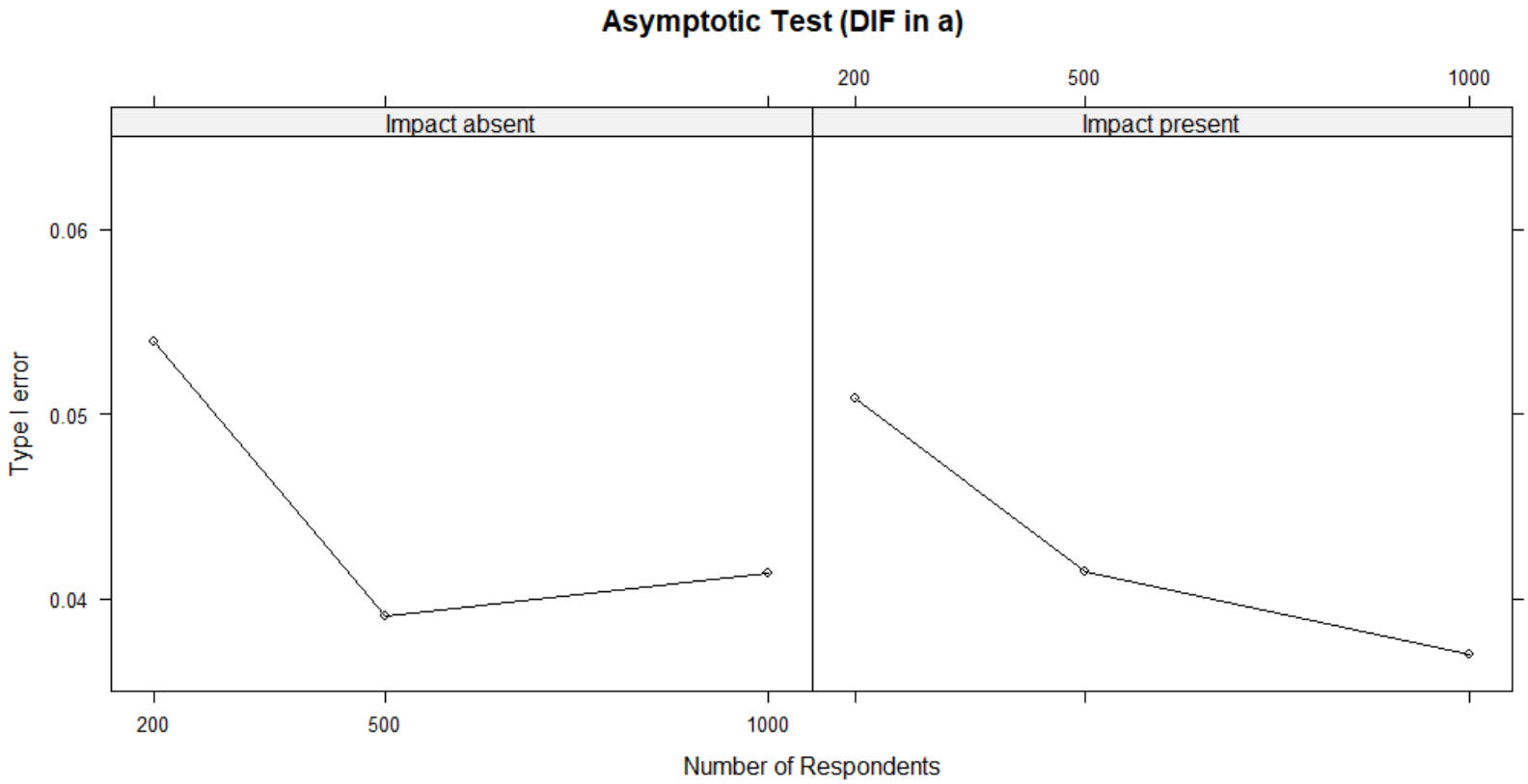
Results for the Bootstrap Test



Results for the Bootstrap Test



Results for the Asymptotic Test



Results for the Asymptotic Test