Sample Size Planning for Latent Variable Models: Classical Approaches and Recent Developments

Rudolf Debelak [1], Felix Zimmer [1]

[1] Universität Zürich

May 18 2023

Overview

Why Sample Size Planning?

Which Latent Variable Models?

Classical Approaches for Power Analysis

Modern Developments

References

Power of Psychological Studies

In the social sciences, we often study the presence of certain effects, for instance:

- Do the means of certain groups in a dependent variable differ?
- Do two variables correlate with each other?

To detect such effects with statistical tests in a frequentist setting, we usually use null hypothesis tests.

Power of Psychological Studies

Power is the probability of detecting an effect present in the population (Button et al., 2013).

- Psychological studies need high power: if the power is low, effects may be missed.
- Studies with very high power usually use large samples and can therefore be (too) expensive.

Therefore, we need to plan the power of our studies (Cumming, 2014).

Power of Psychological Studies

In classical methods for sample size planning, we need to consider at least four variables to determine the desired sample size:

- Desired power
- Significance level
- Effect size
- The statistical test

Power of Psychological Studies

In applied psychological studies, research questions are often infinitely more complex. Consider, for example, the evaluation of an assessment (e.g., an intelligence test):

- Which models can be used in such a scenario?
- What are the effect sizes?

There are many possible answers to both questions. But even if we have answers, we need suitable methods for sample size planning.

Power of Psychological Studies

- In this talk, I will briefly discuss analytical and recently developed simulation-based approaches for carrying out a power analysis with latent variable models.
- I focus on models in the framework of item response theory. They are strongly related to categorical factor analysis models.
- While the presented analytical approaches turn out to be specific for certain models and tests, the later approaches can be generalized well to other types of statistical models.

Statistical Models for Assessments

- Models of item response theory are widely used to describe the interaction of test items (tasks) and respondents (test takers). In assessments, each such response is given by a categorical variable (e.g., $0 =$ task not solved, $1 =$ task solved).
- Structurally, models of item response theory are related to ordinal regression models and latent factor models, but developed in a somewhat independent tradition.
- To provide an idea of the overall setup, I will briefly discuss one such model, the two-parametric logistic (2PL) model of Birnbaum, on the next slide.

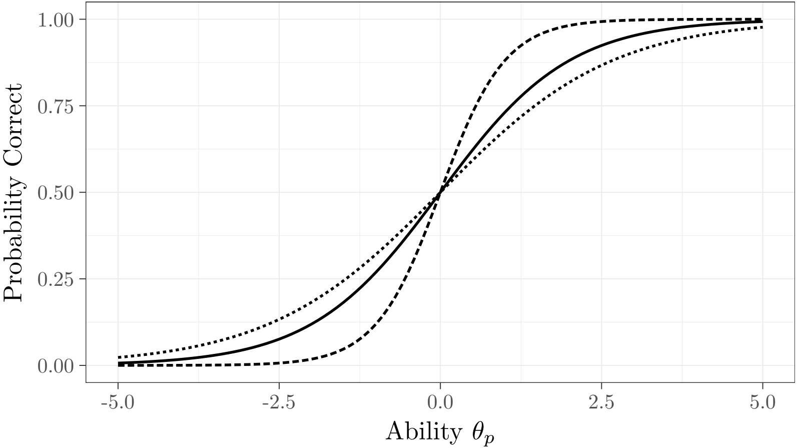
Statistical Models for Assessments

The model equation of the Birnbaum model is:

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

- i is an index for a person, j is an index for an item.
- X_{ij} is the response of person i to item j. This response can be 0 or 1 based on our encoding.
- a_j is a model parameter for item j, which represents the item discrimination.
- b_j is a model parameter for item j, which represents the item difficulty.
- θ_i is a model parameter for person i, which represents the person ability.
- $P(X_{ij} = 1 | a_j, b_j, \theta_i)$ is the probability of answer 1, i.e., a correct answer, given θ_i, a_j and b_j.

Statistical Models for Assessments



Statistical Models for Assessments

Examples for interesting hypotheses:

- Are estimates for a_j stable for different groups of respondents?
- Are estimates for b_j stable for different groups of respondents?
- Can we set equalities between sets of parameters? Can we, for instance, assume that all a_j are identical (which leads to the Rasch model)?
- Do estimates of the parameters in new samples deviate from the estimates in previous studies?

A statistical power analysis is necessary

- to know whether relevant alternative models can be detected, and
- to plan the sample size accordingly.

Classical Approaches for Power Analysis

Analytical approaches for power analysis for item response theory models were developed for some specific cases:

- The M_2 test, an omnibus test for detecting violations of IRT models.
- Testing linear hypotheses based on the likelihood ratio, score, Wald and gradient test (Draxler, 2010; Draxler & Alexandrowicz, 2015; Zimmer, Draxler, & Debelak, 2022).

In this presentation, I will briefly present some work on the second approach.

Classical Approaches for Power Analysis

I will demonstrate one such classical approach using a simple example. First, we need to define the following essential parts:

- Null hypothesis: The 2PL model
- Alternative hypothesis: The 2PL model with different parameters in one item for predefined groups.
- Statistical Test: The likelihood ratio test.

Classical Approaches for Power Analysis

Remember the model equation of the 2PL model:

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

We can formulate our null hypothesis in the form of the following equation:

$$\begin{pmatrix} a_{4A} \\ b_{4A} \end{pmatrix} = \begin{pmatrix} a_{4B} \\ b_{4B} \end{pmatrix}$$

Classical Approaches for Power Analysis

To check this null hypothesis, we compare the relative fit of two models:

- A 2PL model that assumes the item parameters of item 4 to be identical in all groups.
- A 2PL model that allows the item parameters of item 4 to be different in at least two groups.

These lead to different parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

Classical Approaches for Power Analysis

The likelihood ratio statistic now corresponds to:

$$2 \sum_{x \in X} (l_{\hat{\beta}_0}(x) - l_{\hat{\beta}_1}(x)) \hat{h}(x)$$

- X is the set of all possible response vectors.
- $l_{\hat{\beta}_0}$ and $l_{\hat{\beta}_1}$ are the log-likelihood of both models.
- $\hat{h}(x)$ is the relative frequency of response pattern x in the data.

Under the null hypothesis, this test statistic follows an approximate χ^2 distribution, which can be used for hypothesis testing and to carry out a power analysis.

Classical Approaches for Power Analysis

Some asymptotically equivalent hypothesis tests are:

- The score test
- The Wald test
- The gradient test

In finite samples, these tests differ slightly with regard to their power and how close they are to the assumed χ^2 distribution.

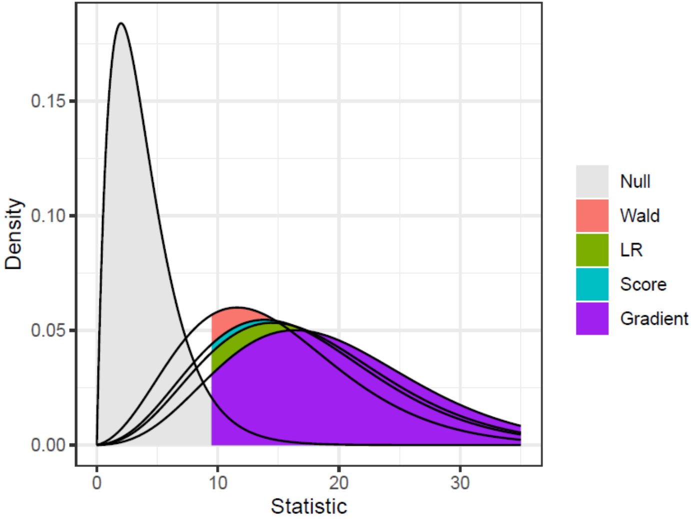
Classical Approaches for Power Analysis

Under the alternative hypothesis, the Wald, LR, Score and gradient test statistics follow asymptotically a non-central χ^2 distribution. For the LR test, the noncentrality parameter λ is calculated by:

$$\lambda = 2n \sum_{x \in X} (l_{\hat{\beta}_0}(x) - l_{\hat{\beta}_1}(x)) \hat{h}_{\beta_1}(x)$$

- X is the set of all possible response vectors.
- $l_{\hat{\beta}_0}$ and $l_{\hat{\beta}_1}$ are the log-likelihood of both models.
- $\hat{h}_{\beta_1}(x)$ is the expected relative frequency of response pattern x in the data under the alternative model.
- n is the sample size.

Classical Approaches for Power Analysis



Classical Approaches for Power Analysis

A disadvantage of this analytical approach is that its calculation requires a summation over all possible response patterns X:

$$2n\sum_{x\in X}(l_{\hat{\beta}_0}(x) - l_{\hat{\beta}_1}(x))\hat{h}_{\beta_1}(x)$$

The number of calculation steps increases exponentially with the number of items. For long tests, Zimmer et al. (2022) proposed the following sampling-based approach:

- Generate one large artificial dataset using the alternative model.
- Calculate the statistic based on the maximum likelihood estimates.
- Retrieve the noncentrality parameter from the calculated test statistic.

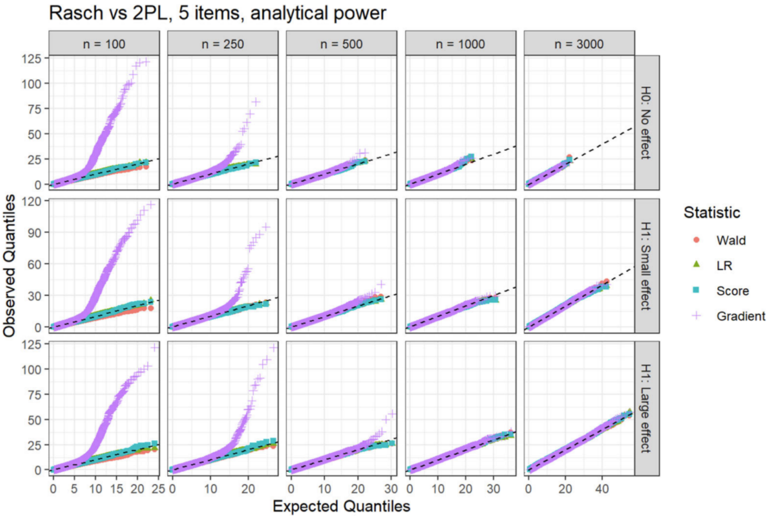
Classical Approaches for Power Analysis

An evaluation with simulation studies:

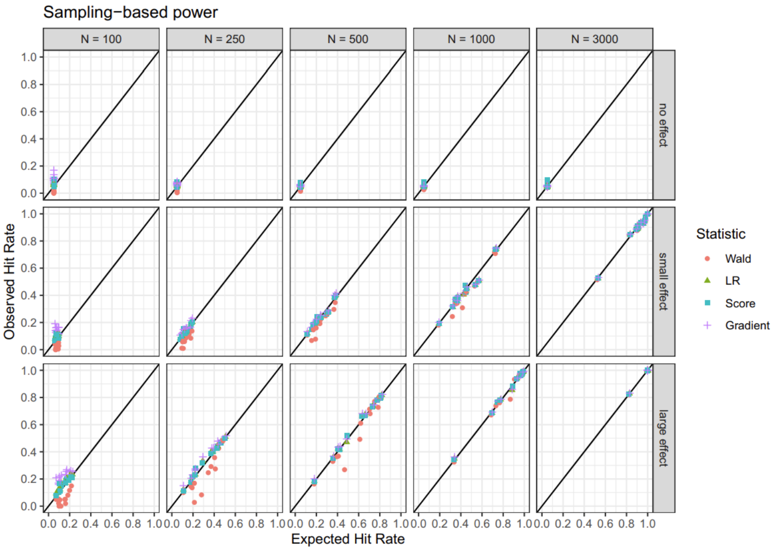
- Two Hypotheses (Differential Item Functioning, Rasch vs 2PL)
- Number of items: 10, 50
- Sample Sizes: 500, 1000, 3000
- Size of Effect: No, Small, Large

For each combination of conditions, we simulated 5000 datasets.

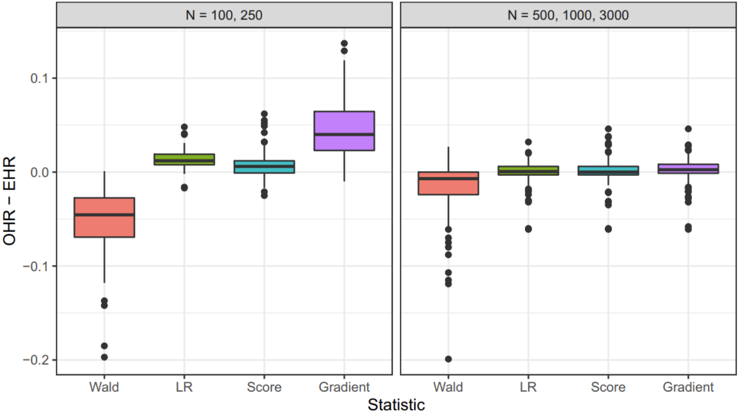
Classical Approaches for Power Analysis



Classical Approaches for Power Analysis



Classical Approaches for Power Analysis



Classical Approaches for Power Analysis

- The outlined approach was found to lead to accurate results for several important item response models and for the four proposed tests (Draxler, 2010; Zimmer et al., 2022).
- It is implemented in the `tcl` (Draxler & Kurz, 2023) and `irtpwr` (Zimmer & Debelak, 2022a) software packages for R.
- Although widely applicable, this approach has several practical limitations: a) It requires that maximum likelihood estimation is possible and available. b) It requires that we are testing two nested models against each other, with the more specific model resulting as linear form from the more general one.

Modern Developments

There are several additional questions that cannot be easily addressed in a classical power analysis:

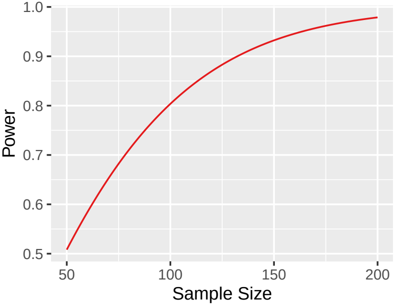
- Sometimes, statistical tests such as the likelihood ratio test are not applicable (e.g., in the modeling of guessing parameters).
- A classical power analysis provides an estimate of the sample size required to detect a specific effect. How can we consider additional design aspects that affect the cost of a study?
- ...

The next approach will try to address these questions using machine learning.

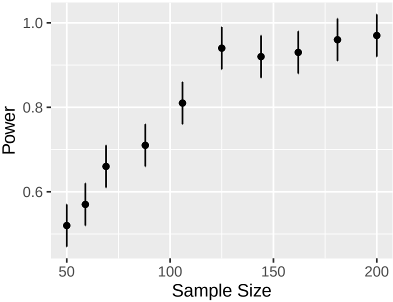
Modern Developments

Usually we can also use simulations to conduct a power analysis:

Analytical Power Analysis



Simulation-based Power Analysis



Modern Developments

In a simulation-based power analysis, power is be determined using simulated data sets:

1. We simulate often (for example 1000x) data sets with a) a certain sample size and b) with the effect size we assume.
2. We choose c) a statistical test with d) a significance level.
3. We determine in each simulated data set how often our test becomes significant. The proportion of significant results in all data sets then estimates the power.

Modern Developments

Simulation-based power analysis is flexible and intuitively clear, but it has some practical shortcomings.

- We might have to code parts of the evaluation, such as the statistical test.
- We might have to code parts of the model generation.
- We have to make a decision for which sample sizes we want to evaluate the power.
- We have to make a decision what kind of samples we want to simulate. Several different kinds of samples can correspond to the same effect size (e.g., differential item functioning).

Modern Developments

In the proposed approach, we first define a **data generating function** (DGF), based on an effect size, a significance level, and a chosen statistical test.

- Input: properties of the study design, e.g., a given sample size.
- Output: We apply the defined statistical test to an artificial dataset with the chosen design and effect size. How often is the result significant?

This leads to an estimate of the power for a given design and a given sample size, including a confidence interval.

Modern Developments

- We now know the power for certain study designs, e.g., for certain sample sizes.
- We can now use this information to extrapolate to designs for which the power was not yet evaluated.
- For this extrapolation, we apply machine learning methods.

Modern Developments

In summary, we propose the following overall procedure: In an initialization step, we first determine the power for different study designs via simulation. We then we alternate the following steps until a stop criterion is met:

- We try to model the relationship between design and power via a machine learning model which serves as surrogate model.
- These models provide us with points where the target power is achieved. We then check the actual power at these points, and update our models.

Modern Developments

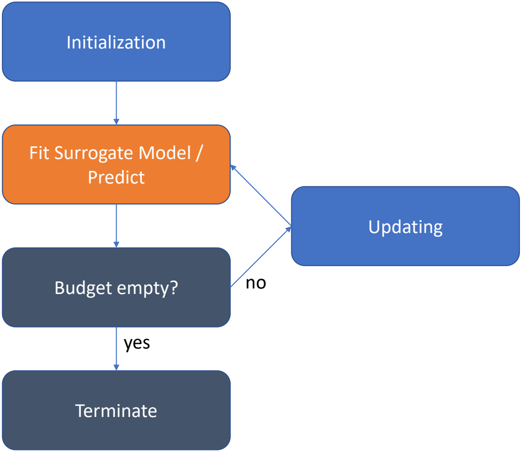


Figure: **The Surrogate Modeling Approach**

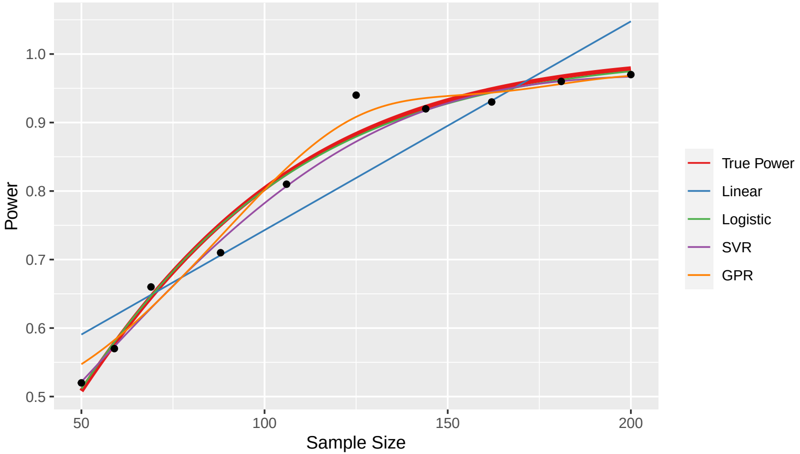
Modern Developments

Initialization: We start with a few study designs for which we evaluate power via simulation. This initial evaluation requires the following decisions:

- We set upper and lower bound for the considered design parameters (e.g., sample size).
- We define how many designs we want to consider.
- We define the computational budget for the initial evaluation (e.g., 10 % of the total budget).

For example: We select 10 sample sizes between 10 and 1000, and generate 1000 simulated datasets to evaluate the power for each size.

Modern Developments



Modern Developments

After the initialization, the surrogate model provides a candidate design that is likely to achieve the target power. We further evaluate the power for this design and update the surrogate model. We stop when a stop criterion is met, e.g.

- our computational budget is empty **or**
- we have achieved the target power sufficiently well.

These methods yield designs that lead to a desired target power. In a second step, we can evaluate the costs of these designs to find study designs with minimal costs.

Modern Developments

We evaluated these ideas with simulated data, with a focus on problems from psychology.

We addressed two sets of problems:

1. Mincost: Which designs with a desired level of power lead to minimal costs?
2. (Maxpower: Which designs with a desired level of costs lead to maximum power?)

We used the following surrogate models:

- Gaussian Process Regression
- Support Vector Machines
- Linear Regression (only one-dimensional)
- Logistic Regression (only one-dimensional)

Modern Developments

Investigated problems:

1. Student t-test for mean comparisons of two groups of the same size and normally distributed values.
2. ANOVA for mean comparisons of k groups of the same size.
3. Student t-test for mean comparisons of two groups of the same size and skewed distributed values.
4. **LRT for comparison of Rasch and 2PL model in item response theory.**
5. Generalized mixed model: determination of a fixed effect with 3 groups of the same size.
6. Generalized mixed model: determination of a fixed effect with k groups of the same size.

We varied the sample size n and k in problems 2 and 6, respectively.

Modern Developments

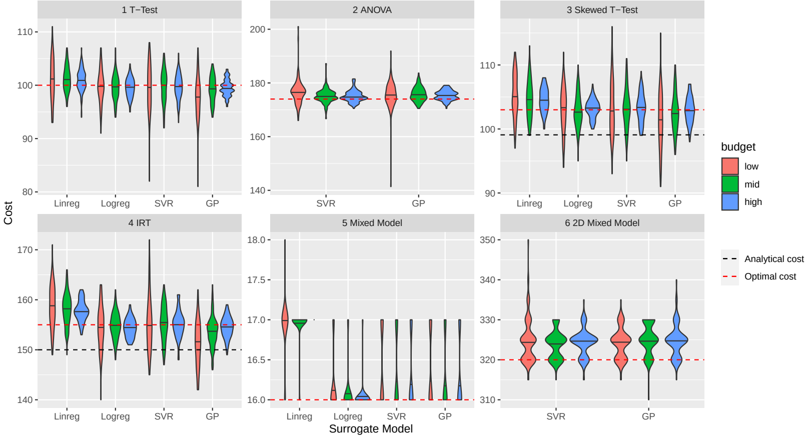
Each problem was combined with three budget sizes, i.e. number of simulated datasets:

- 1000, 2000, and 4000 for problems with no group parameter k.
- 2000, 4000, and 8000 for problems with group parameter k.

For each condition, the analysis was repeated 100 times to check the stability of the outcomes. The results were compared with analytical power analysis whenever possible.

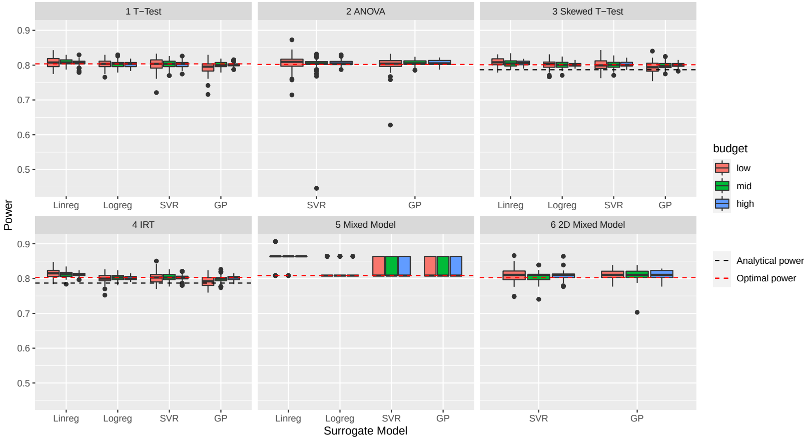
Modern Developments

Mincost: total cost



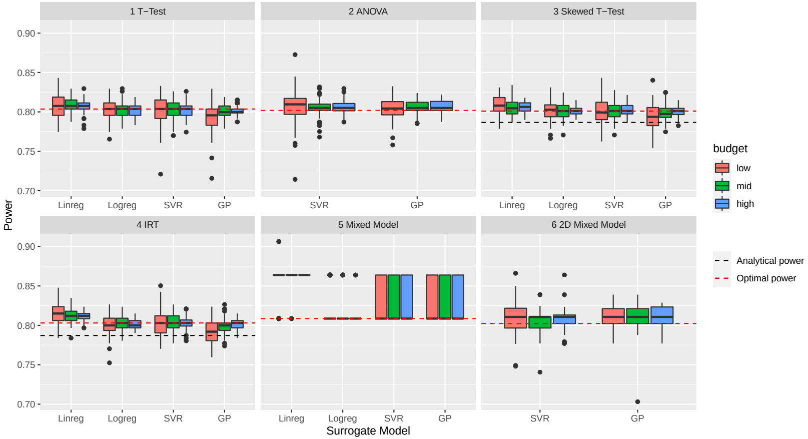
Modern Developments

Mincost: Power (with outlier)



Modern Developments

Mincost: Power (without outliers)



Modern Developments

- For additional results, see our preprints: https://psyarxiv.com/tnhb2/ and https://psyarxiv.com/r9w6t/
- The presented machine-learning based approach for study planning is implemented in the R package `mlpwr` (Zimmer & Debelak, 2022b).
- Compared to the analytical approaches that were presented first, this approach is computationally more intensive, but more flexible with regard to the definition of the tests and models for which the sample size is planned.
- The work on the machine-learning based approach is still ongoing.

References I

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376. doi: 10.1038/nrn3475

Cumming, G. (2014). The new statistics: why and how. *Psychological Science*, *25*(1), 7–29. doi: 10.1177/0956797613504966

Draxler, C. (2010). Sample size determination for Rasch model tests. *Psychometrika*, *75*(4), 708–724. doi: 10.1007/s11336-010-9182-4

Draxler, C., & Alexandrowicz, R. W. (2015). Sample size determination within the scope of conditional maximum likelihood estimation with special focus on testing the Rasch model. *Psychometrika*, *80*(4), 897–919. doi: 10.1007/s11336-015-9472-y

References II

Draxler, C., & Kurz, A. (2023). tcl: Testing in conditional likelihood context [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tcl (R package version 0.2.0)

Zimmer, F., & Debelak, R. (2022a). irtpwr: Power analysis for irt models using the wald, lr, score, and gradient statistics [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=irtpwr (R package version 1.0.0)

Zimmer, F., & Debelak, R. (2022b). mlpwr: A power analysis toolbox to find cost-efficient study designs [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=mlpwr (R package version 1.0.0)

References III

Zimmer, F., Draxler, C., & Debelak, R. (2022). Power analysis for the wald, lr, score, and gradient tests in a marginal maximum likelihood framework: Applications in irt. *Psychometrika*. doi: https://doi.org/10.1007/s11336-022-09883-5