Using psychotools for Evaluating the Performance of Score-Based Measurement Invariance Tests in IRT Models

Lennart Schneider
Ludwig-Maximilians-Universität München

Carolin Strobl
Universität Zürich

Achim Zeileis
Universität Innsbruck

Rudolf Debelak
Universität Zürich

Abstract

The detection of differential item functioning (DIF) is a central topic in psychometrics and educational measurement. In the past few years, a new family of score-based tests of measurement invariance has been proposed, which allows the detection of DIF along arbitrary person covariates in a variety of item response theory (IRT) models. Schneider, Strobl, Zeileis, and Debelak (2020) illustrate the application of these tests within the R system for statistical computing. This vignette targets more advanced users and provides a tutorial on how to conduct simulation studies investigating the performance of score-based tests of measurement invariance.

Keywords: Item response theory, IRT, score-based tests, measurement invariance, differential item functioning, DIF, structural change.

Schneider *et al.* (2020) describe the conceptual framework as well as the software to perform differential item functioning (DIF) investigations. This vignette aims at more advanced users and provides a tutorial on how to conduct simulation studies investigating the performance of score-based tests of measurement invariance. We recommend to have read at least the following sections of Schneider *et al.* (2020): "A Conceptual and Formal Framework for Score-Based Measurement Invariance Tests" and "The Implementation of Score-Based Measurement Invariance Tests within R".

In this vignette, we want to carry out simulation studies that investigate how to appropriately model impact with a 2PL model in the presence of a continuous numerical covariate (Simulation 1) and what power we can expect to detect DIF in this case (Simulation 2). The motivation for the first simulation study is that modeling impact is necessary to avoid an increased Type I error rate, as was already mentioned in the main text, whereas the second simulation study aims at providing an additional investigation of the method's power.

Both of these simulations serve as illustrations. To keep our presentation concise, the design of these studies is somewhat limited. To provide a realistic setting, we again rely on the first six items of the Verbal Aggression dataset, which were already used in the "Illustrations with Empirical Data" section of Schneider *et al.* (2020). In the following, `refmodel` refers to a 2PL model fitted to the first six items of the Verbal Aggression dataset:

```
refmodel <- nplmodel(VerbalAggression$resp2[, 1:6])
```

Instead of using the item parameters of this fitted 2PL model, we could of course also simply

generate them based on specific parametric distributions, such as a standard normal distribution for the item difficulty parameters. This could be done using standard functions of R, which we do not present for brevity.

In our first simulation study, we investigate the Type I error for our score-based tests when applied to a 2PL model. Here, no DIF is present, but we use various conditions that differ with regard to the presence of impact, the number of groups used to model it, and the relationship of the covariate with the person parameters. We are interested how different methods of modeling impact affect the Type I error rate depending on the relationship between the ability of respondents and the observed covariate. In summary, we want to vary the following conditions in our simulation study:

- The presence of impact. The person parameter distribution could be either normal ($N(0, 1)$) for the whole sample, or a mixed normal distribution. In the second case, there are two latent groups of respondents, whose person parameters follow a $N(-0.5, 1)$ or $N(0.5, 1)$ distribution, respectively. These two cases correspond to conditions without and with impact.

- The type of relationship between the ability parameter θ_i and the observed covariate cov_i, considered over all respondents $i = 1, \ldots, N$. In a first condition, there is no systematic relationship, and the covariate is generated independently from the ability parameter. In a second condition, there is a linear relationship. Using an error term ε, which follows a standard normal distribution, this is denoted by $\text{cov}_i = \theta_i + \varepsilon_i$. Finally, we consider a quadratic relationship $\text{cov}_i = \theta_i^2 + \varepsilon_i$ as a third condition.

- We further vary the number of groups G which are used for modeling impact, using 1, 2, 5 and 25 groups. For simplicity, we assume that impact should always be modeled based on groups of about equal size that correspond to respondents whose value in the covariate come from different intervals. The boundaries of these intervals are therefore defined based on percentiles of the observed distribution of the covariates.

We hypothesize that an independent relationship should show no systematic effect on the Type I error rate, and that modeling a quadratic relationship should be more difficult than a linear one and thus require a larger number of groups. An heuristic argument for this expectation is that in a quadratic relationship, the change of the expected personality parameter becomes very large for a comparatively small group of respondents (namely those with a very high or very low covariate). It seems plausible that the resulting model is more difficult to estimate than a model resulting from a linear relationship. On the other hand, we keep the following conditions fixed:

- The sample size `N` is 1000.

- The used item parameters correspond to the item parameter estimates for the six items of the verbal aggression dataset.

Our data generating process (DGP) thus consists of the following steps:

- Generating the vector of person parameters (`theta`) for `N` persons following a normal distribution (standard normal if no impact is present, `impact = FALSE`, and $N(-0.5, 1)$ or $N(0.5, 1)$ for each half of the sample if impact is present, `impact = TRUE`).

- Generating the vector of covariates for `N` persons, either independent, in a linear relationship or in a quadratic relationship.
- Determining the number of groups, `G`, for modeling impact and modeling the impact effect if the number of groups, `G`, is larger than one (i.e., categorizing the covariate based on equidistant percentiles matching the number of groups, using `cut`; see also `?cut`).
- Simulating data under the 2PL model using the item parameters of our already fitted model (using the `rpl` function of the *psychotools* package, see `?rpl` for more information).

Listing 1 shows the corresponding code.

Listing 1: The data generating process

```
dgp <- function(model, N = 1000, G = 1, impact = FALSE, cotype = "random") {

  mu <- if(impact) rep_len(c(-0.5, 0.5), N) else rep_len(0, N)
  theta <- rnorm(N, mu, 1)

  covariate <- switch(as.character(cotype),
    "random" = rnorm(N, 0, 1),
    "linear" = theta + rnorm(N, 0, 1),
    "quadratic" = theta ^ 2 + rnorm(N, 0, 1)
  )

  d <- data.frame(theta = theta, covariate = covariate)
  if(G > 1) {
    d$impact <- cut(covariate, labels = 1:G, include.lowest = TRUE,
      breaks = quantile(covariate, probs = 0:G / G))
  }

  d$resp <- rpl(theta,
    a = discrpar(model),
    b = itempar(model),
    g = guesspar(model),
    u = upperpar(model),
    return_setting = FALSE)

  return(d)
}
```

To calculate p-values three steps are needed: Simulate data (`dgp(...)`), fit the 2PL model (`nplmodel(...)`) and calculate the score-based tests (`sctest(...)`). A possible solution is given in Listing 2 using, e.g., 1000 persons (`N = 1000`), one group (`G = 1`), simulating no impact (`impact = FALSE`) and assuming the covariate to be independent of the person parameters (`cotype = "random"`).

Listing 2: Calculate p-values

```
d <- dgp(refmodel, N = 1000, G = 1, impact = FALSE, cotype = "random")
m <- nplmodel(d$resp, impact = d$impact, vcov = FALSE)
sctest(m, order.by = d$covariate, functional = "DM")
```

In our simulation, these three steps are repeated `M` times under all simulated conditions. We are interested in the hit rate under each condition, which is calculated as the proportion of significant tests given a specified significance level `alpha`. As no DIF effect was simulated, this is the Type I error. In the following code, we also allow for setting the `parm` argument that allows for only testing a specified subset of the item parameters; `parm = NULL` defaults

to using all item parameters. See Listing 3 for the code using, e.g., `M = 1000` replications, and setting `alpha` to `0.05`.

Listing 3: Calculate hit rate

```
hitrate <- function(model, M = 1000, alpha = 0.05, parm = NULL, N = 1000,
  G = 1, impact = FALSE, cotype = "random") {

  pval <- replicate(M, {
    d <- dgp(model, N = N, G = G, impact = impact, cotype = cotype)
    m <- nplmodel(d$resp, type = "2PL", impact = d$impact,
      maxit = 5000, reltol = 1e-4, vcov = FALSE)
    sctest(m, order.by = d$covariate, functional = "DM", parm = parm)$p.value
  })
  mean(pval < alpha)
}
```

The Type I error is investigated for a varying number of groups (`G = c(1, 2, 5, 25)`) that are used to model impact (which can be present or not, `impact = c(FALSE, TRUE)`) and the different types of relationship of the covariate with the person parameters, `cotype = c("random", "linear", "quadratic")`. Listing 4 shows the final code.

Listing 4: Simulation

```
sim <- function(model, M = 1000, alpha = 0.05, parm = NULL, N = 1000,
  G = c(1, 2, 5, 25), impact = c(FALSE, TRUE),
  cotype = c("random", "linear", "quadratic")) {

  d <- expand.grid(G = G, impact = impact, cotype = cotype)
  d$hitrate <- NA
  for(i in seq_len(NROW(d))) {
    d$hitrate[i] <- hitrate(model, M = M, alpha = alpha, parm = parm, N = N,
    G = d$G[i], impact = d$impact[i], cotype = d$cotype[i])
  }
  return(d)
```

Results are given in Table 1. As expected, a random relationship of the covariate with the person parameters does not require impact modeling in any case, i.e., we observe Type I error rates close to 5% for the single group model regardless of whether impact is present or not, and more conservative Type I error rates when using multiple groups. Looking at the linear relationship, we see that a single group model fails to yield reasonable Type I error rates, but we achieve rates around the nominal 5%, using five groups. Finally, the hardest case of a quadratic relationship would require more than 25 groups if impact is present to achieve a Type I error rate close to 5%. Nevertheless, we can observe the trend that if the number of groups grows, the Type I error rate is closer to its nominal level.

In a second simulation study, we investigate the power of a 2PL model assuming uniform DIF in the first item. We use the (somewhat arbitrary) condition that the item difficulty parameter of this item is changed by *sd* for persons exhibiting a covariate larger than the median, making it more difficult for these respondents, but is unchanged for the remaining sample. *sd* is simply one standard deviation of all item difficulty parameters. We can reuse almost all of the code presented above in Simulation 1. However, we do have to add DIF, i.e., the last part of our DGP now looks like the following:

Listing 5: The data generating process when simulating DIF

```
  itempar_dif <- itempar(model)
  itempar_dif[1] <- itempar_dif[1] + sd(itempar_dif)
```

Table 1: Simulation 1: Results on the Type I error in the 2PL model.

Relationship	Impact	Type I Error (%)			
		Single Group	No. Groups		
			2	5	25
Random	No	4.60	1.60	1.40	1.00
	Yes	4.80	1.70	2.10	1.30
Linear	No	100.00	96.80	4.00	1.50
	Yes	100.00	99.00	5.40	1.40
Quadratic	No	100.00	98.80	47.40	6.60
	Yes	100.00	99.60	75.50	10.70

```
19
20  dif_id <- covariate > median(covariate)
21
22  d$resp <- matrix(NA, N, length(model$items))
23  d$resp[!dif_id, ] <-
24    rpl(theta[!dif_id],
25      a = discrpar(model),
26      b = itempar(model),
27      g = guesspar(model),
28      u = upperpar(model),
29      return_setting = FALSE)
30  d$resp[dif_id, ] <-
31    rpl(theta[dif_id],
32      a = discrpar(model),
33      b = itempar_dif,
34      g = guesspar(model),
35      u = upperpar(model),
36      return_setting = FALSE)
```

This is the only necessary change. Since we have included a model violation, our hit rate now represents the power. Results are given in Table 2. To evaluate our findings, we have to consider which scenarios yielded a reasonable Type I error rate close to 5% in our first study. Looking at the random relationship of the covariate with the person parameters, we observe a high power in all scenarios, with impact being present resulting in a slightly lower power. Regarding the linear relationship, we observe a power of around 9% to 15% for the scenarios that yielded a reasonable Type I error rate beforehand. Detecting uniform DIF in the first item being one standard deviation more difficult for persons exhibiting a covariate larger than the median appears to be especially difficult if the relationship of the covariate with the person parameters is linear. A possible explanation is that, if both the impact and the DIF effect are linearly related to the person parameters, modeling the impact effect can essentially mask a part of the DIF effect. Finally, in a scenario with a quadratic relationship, the scenario of no impact being present and a multiple group model using 25 groups results in a high power of around 86%. If impact is present, the power is also high (at around 83%). However, we have to keep in mind that the Type I error rate was already at around 10% in this scenario; that is, we would have observed an increased rate of significant results even if no DIF is present.

The full simulation code of both simulations can be inspected using `demo("toolbox1", package = "psychotools")` or `demo("toolbox2", package = "psychotools")`.

All results were obtained using the R system for statistical computing R Core Team (2021) version 3.5.3 employing the add-on packages *mirt* Chalmers (2012) version 1.31, *psychotools*

Table 2: Simulation 2: Results on the power to detect DIF in the first item using the 2PL model.

Relationship	Impact	Power (%)			
		Single Group	No. Groups		
			2	5	25
Random	No	94.60	90.40	91.40	89.80
	Yes	94.10	88.20	89.40	87.40
Linear	No	100.00	98.20	15.20	12.30
	Yes	100.00	99.80	11.50	8.80
Quadratic	No	100.00	100.00	95.50	86.30
	Yes	100.00	100.00	96.60	83.10

Zeileis, Strobl, Wickelmaier, Komboz, Kopf, Schneider, and Debelak (2021) version 0.6-0 and *strucchange* Zeileis, Leisch, Hornik, and Kleiber (2002) version 1.5-2, which are freely available under the General Public License from the Comprehensive R Archive Network at https://cran.r-project.org/. Numerical values were rounded based on the IEC 60559 standard.

References

Chalmers RP (2012). "**mirt**: A Multidimensional Item Response Theory Package for the R Environment." *Journal of Statistical Software*, **48**(6), 1–29. doi:10.18637/jss.v048.i06.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schneider L, Strobl C, Zeileis A, Debelak R (2020). "An R Toolbox for Score-Based Measurement Invariance Tests in IRT Models." *PsyArXiv r9w34*, PsyArXiv Preprints. doi:10.31234/osf.io/r9w34.

Zeileis A, Leisch F, Hornik K, Kleiber C (2002). "**strucchange**: An R Package for Testing for Structural Change in Linear Regression Models." *Journal of Statistical Software*, **7**(2), 1–38. doi:10.18637/jss.v007.i02.

Zeileis A, Strobl C, Wickelmaier F, Komboz B, Kopf J, Schneider L, Debelak R (2021). ***psychotools***: *Psychometric Modeling Infrastructure*. R package version 0.6-0, URL https://CRAN.R-project.org/package=psychotools.

Affiliation:

Lennart Schneider
Department of Statistics
Ludwig-Maximilians-Universität München
Ludwigstraße 33
80539 München, Germany
E-mail: lennart.sch@web.de

Carolin Strobl, Rudolf Debelak
Department of Psychology
Universität Zürich
Binzmühlestr. 14
8050 Zürich, Switzerland
E-mail: Carolin.Strobl@psychologie.uzh.ch, Rudolf.Debelak@psychologie.uzh.ch
URL: http://www.psychologie.uzh.ch/fachrichtungen/methoden.html

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: https://www.zeileis.org/