Item-Wise DIF Detection Using ScDIFtest: An Illustration With the SPISA data

Dries Debeer[1, 2] and Valentina Tomasulo[2]

[1] *ITEC, imec research group at KULeuven*
[2] *Univeristy of Zurich*

Abstract

This vignette, explains the installation of the `scDIFtest` package and provides an illustration of item-wise DIF-detection with the `scDIFtest`-function using a subset of the SPISA data set.

The score-based test framework for parameter instability has been proposed for testing measurement invariance in measurement models. Until now, the focus was on (a) testing the invariance of all parameters simultaniously, or (b) on testing the invariance of a single parameter in the model. However in educational and psychological assessments, the appropriatness of each items is of interest. For instance, the detection of differential item function (DIF) plays an important role in validating new items. The `scDIFtest` package provides a user-friendly method for detecting DIF by automatically and efficiently applying the tests from the score-based test framework to the individual items in the assessment. The main function of the `scDIFtest` package is the `scDIFtest` function, which is a wrapper around the `strucchange::sctest`-function.

To detect DIF with the `scDIFtest` package, first, the appropriate Item Response Theory (IRT) or Factor Analysis (FA) model should fitted using the `mirt` package. The `scDIFtest`-function can directly be used on the resulting `mirt`-object. Hence, in addition to the `scDIFtest`, the package `mirt` will typically also be loaded in the `R` session. For now, `scDIFtest` only works for IRT/FA models that were fitted using the `mirt` package, but we aim to extend this to other packages that fit IRT/FA models using maximum likelihood estimation.

Overview of the method

In order to fit the IRT model and analyse DIF with the `scDIFtest`, the following steps are necessary:

1. installation of the `R`-package(s)
2. data preparation
3. fitting the IRT Model by using either the `mirt` or `multipleGroup` function implemented in the `mirt` package (Chalmers, 2012)
4. detecting DIF by using `scDIFtest` (Debeer, 2020)

5. interpreting the results

In the sections that follow, these steps will be explained in detail.

1 Installation

The `scDIFtest` package is installed using the following commands:

```
> install.packages("devtools")
> devtools::install_github("ddebeer/scDIFtest")
```

Since, the `mirt` package (Chalmers, 2012) is required for fitting the IRT/FA model of interest, it should also be installed (using `install.packages("mirt")`).

2 Data preparation

In this vignette, a subset of the SPISA data is used. This data is part of the `psychotree` package, it can be accessed when the `psychotree` package is installed. To load the SPISA dataset:

```
> install.packages("psychotree", quiet = TRUE)
> data("SPISA", package = "psychotree")
```

The SPISA data is a subsample from the general knowledge quiz "Studentenpisa" conducted online by the German weekly news magazine SPIEGEL (Trepte & Verbeet, 2010). The data contain the quiz results from 45 questions as well as sociodemographic data for 1075 university students from Bavaria (Trepte & Verbeet, 2010). Although there were 45 questions addressing different topics, this illustration is limited to the analysis of the nine sience questions (items 37 - 45). To analyze the data with `mirt`, the responses are converted to a data frame.

```
> resp <- as.data.frame(SPISA$spisa[,37:45])
```

In addition to the responses, the SPSA data also contains five sociodemographic variables (i.e., person covariates):

```
> summary(SPISA[,2:6])

    gender          age           semester    elite            spon     
 female:417   Min.   :18.0   2      :173   no :836   never    :303  
 male  :658   1st Qu.:21.0   4      :123   yes:239   <1/month :127  
              Median :23.0   6      :116             1-3/month:107  
              Mean   :23.1   1      :105             1/week   : 79  
              3rd Qu.:25.0   5      : 99             2-3/week : 73  
              Max.   :40.0   3      : 98             4-5/week : 60  
                             (Other):361             daily    :326  
```

In this illustration, we will try to detect DIF along the following three covariates:

- *age* of the student in years (numeric covariate)
- *gender* of the student (unordered categorical covariate)
- and *spon*, which is the frequency of assessing the SPIEGEL online (SPON) magazine (ordered categorical covariate)

3 Fitting the IRT model using either the `mirt` or `multipleGroup` function

It is important to note that, for the package to work, the parameters in the assumed IRT model need to be be estimated using either the `mirt` or the `multipleGroup` function from the `mirt` package. The `multipleGroup` function can model impact between groups of persons, which is not possible with the `mirt` function. Modeling impact is important when the goal is to detect DIF (DeMars, 2010). In this illustration, for instance, we test whether there is impact with respect to gender by comparing a model which allows ability differences between male and female students with a model that assumes there are no group differencesin ability. The relative fit of these two models is compared, and the best fitting model is selected for the DIF analysis. The general idea is that we want to avoid (a) false DIF detections that can be attributed to ability differences and (b) not detecting DIF that is masked by unmodeled ability differences.

First the `mirt` package is loaded in the `R` session:

```
> library(mirt, quietly = TRUE)
```

Then the two models are fit and compared. Note that in general we do not recommend using `verbose = FALSE`, but for this vignette it is more convenient.

```
> fit_2PL <- mirt(data = resp,
                  model = 1,
                  itemtype = "2PL",
                  verbose = FALSE)
> fit_multiGroup <- multipleGroup(
    data = resp, model = 1,
    group = SPISA$gender,
    invariance = c("free_means",
                   "slopes",
                   "intercepts",
                   "free_var"),
    verbose = FALSE)
```

The comparison of the two models with `anova` yields the following results:

```
> anova(fit_2PL, fit_multiGroup)
```

```
Model 1: mirt(data = resp, model = 1, itemtype = "2PL", verbose = FALSE)
Model 2: multipleGroup(data = resp, model = 1, group = SPISA$gender,
    invariance = c("free_means", "slopes", "intercepts", "free_var"),
    verbose = FALSE)

       AIC     AICc    SABIC       HQ      BIC    logLik     X2  df   p
1 10161.68 10162.33 10194.16 10195.64 10251.33 -5062.843    NaN NaN NaN
2 10139.62 10140.41 10175.69 10177.34 10239.22 -5049.808 26.069   2   0
```

The `multipleGroup` model with ability differences between male and female test takers best fits the data (lower AIC and BIC; small *p*-value for the Likelihood Ratio Test). It seem like there are differences between male and female stundents with respect to the assessed science knowledge. Therefore, `multipleGroup` model model is used in the DIF detection analysis.

4 Detecting DIF by using scDIFtest

In the (sub)sections that follow, DIF is tested for three different covariates: gender, age and spon but only the DIF analysis for gender is explained in more detail. Yet the the used `R` commands are the same for any covariate. The interpretation is given for all of the covariates.

4.1 DIF by `gender`

To test itemwise DIF along gender, the `scDIFtest` function is used with the fitted model object and `gender` as the `DIF_covariate` argument. Note that the `scDIFtest` package has to be loaded first.

```
> library(scDIFtest)
> DIF_gender <- scDIFtest(fit_multiGroup, DIF_covariate = SPISA$gender)
```

The resulting object is assigned to `DIF_gender`. For a readable version of the results The `print` method is available. In addition, the `summary` method returns a summary of the results as a data frame.

5 Interpreting the results

In the two subsections that follow, the results regarding the analyses of itemwise DIF by `gender`, `age` and `spon` will be interpreted.

5.1 DIF by `gender`

For the gender covariate, the print method gives the following results:

```
> DIF_gender
```

```
         Score Based DIF-tests for 9 items
         Person covariate: SPISA$gender
         Test statistic type: Lagrange Multiplier Test for Unordered Groups

   item_type n_est_pars       stat      p_value        p_fdr
V1       2PL          2  0.4141020 8.129782e-01 9.146005e-01
V2       2PL          2  8.3162505 1.563685e-02 4.691054e-02
V3       2PL          2  4.8449033 8.870388e-02 1.995837e-01
V4       2PL          2 32.7335352 7.798358e-08 7.018522e-07
V5       2PL          2  3.2679379 1.951535e-01 3.512763e-01
V6       2PL          2  0.4159221 8.122387e-01 9.146005e-01
V7       2PL          2 30.3499936 2.567927e-07 1.155567e-06
V8       2PL          2  0.1517182 9.269468e-01 9.269468e-01
V9       2PL          2  0.5925442 7.435851e-01 9.146005e-01
```

First, in three lines some general information is given:

- the type of test that is performed
- the covariate along which DIF is tested (in this case `gender`) and
- the test statistic which is used, in this case the Lagrange-Multiplier-Test for unordered covariates, (LMUO; Merkle & Zeileis, 2013; Merkle, Fan, & Zeileis, 2014).

After these three lines, a table with the main results is printed with one line for each item that was included in the DIF detection analysis. The columns of the table represent:

- the name of each item (in this case "V1" - "V9")
- `item_type`: the type of IRT model used for each item (in this case the two-Parameter Logistic Model (2PL))
- `n_est_pars`: the number of estimated parameters for each item
- `statistic`: the value for the statistic per item (in this case the LMuo statistic)
- `p-value`: the *p*-value per item
- `p.fdr`: the False-Discovery-Rate corrected *p*-value (Benjamini & Hochberg, 1995)

The printed output indicates that, when a significance level of .05 is used, DIF along `gender` is detected in item V4 and in item V7: these two items function differently, depending on the gender of the students.

When one of more items are selected using the `item_selection` argument of the `print` method, the underlying `sctest` objects (or M-fluctuation tests) are printed.

```
> print(DIF_gender, item_selection = c("V4", "V7"))
```

```
        DIF-test for V4
        Person covariate: SPISA$gender
        Test statistic type: Lagrange Multiplier Test for Unordered Groups

        M-fluctuation test

        data:  resp
        f(efp) = 32.734, p-value = 7.798e-08

        DIF-test for V7
        Person covariate: SPISA$gender
        Test statistic type: Lagrange Multiplier Test for Unordered Groups

        M-fluctuation test

        data:  resp
        f(efp) = 30.35, p-value = 2.568e-07
```

Note that here the uncorrected p-values are given.

5.2 DIF by age

The results for the DIF-detection analysis with `age` as the covariate are:

```
> DIF_age <- scDIFtest(fit_multiGroup, DIF_covariate = SPISA$age)
> DIF_age

        Score Based DIF-tests for 9 items
        Person covariate: SPISA$age
        Test statistic type: Double Maximum Test

   item_type n_est_pars      stat     p_value      p_fdr
V1       2PL          2 1.0593393 0.378630317 0.56794548
V2       2PL          2 0.7508117 0.859974883 0.96747174
V3       2PL          2 1.3579887 0.097556732 0.21950265
V4       2PL          2 1.6092879 0.022393893 0.06718168
V5       2PL          2 1.0936080 0.332120746 0.56794548
V6       2PL          2 1.6830445 0.013808746 0.06213936
V7       2PL          2 0.5720489 0.989797256 0.98979726
V8       2PL          2 0.7729229 0.830878151 0.96747174
V9       2PL          2 1.9126378 0.002656523 0.02390871
```

In this case, the Double Maximum Test for continuous numeric orderings (dm; Merkle & Zeileis, 2013; Merkle et al., 2014) is used. The results indicate that DIF along `age` is detected in three items: V4 ($p = 0.022$), V6 ($p = 0.014$), and V9 ($p = 0.003$). Note that the score-based framework has the power to detect DIF along numeric covariates, without assuming some functional form of the DIF.

5.3 DIF by spon

The results for the DIF-detection analysis with spon as the covariate are:

```
> DIF_spon <- scDIFtest(fit_multiGroup, DIF_covariate = SPISA$spon)
> DIF_spon

          Score Based DIF-tests for 9 items
          Person covariate: SPISA$spon
          Test statistic type: Maximum Lagrange Multiplier Test for Ordered
          Groups

   item_type n_est_pars     stat    p_value     p_fdr
V1       2PL          2 1.868941 0.77948065 0.8769157
V2       2PL          2 6.342694 0.13845728 0.4436186
V3       2PL          2 2.390256 0.66388944 0.8535721
V4       2PL          2 3.597938 0.43203919 0.6480588
V5       2PL          2 7.536444 0.08132661 0.4436186
V6       2PL          2 4.847357 0.26199955 0.5319802
V7       2PL          2 1.304980 0.89486619 0.8948662
V8       2PL          2 6.174822 0.14787287 0.4436186
V9       2PL          2 4.553582 0.29554456 0.5319802
```

In this case, the maximum Lagrange-Multiplier-Test (maxLMo; Merkle & Zeileis, 2013; Merkle et al., 2014) is used. Since all tests result in large p-values, we conclude that no DIF was detected along the spon covariate.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. Retrieved from http://www.jstor.org/stable/2346101

Chalmers, R. P. (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

Debeer, D. (2020). scdiftest: Item-wise score-based dif tests [Computer software manual]. (R package version 0.1.0-01)

DeMars, C. E. (2010). Type i error inflation for detecting dif in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961-972. doi: 10.1177/0013164410366691

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.

Trepte, S., & Verbeet, M. (Eds.). (2010). *Allgemeinbildung in deutschland - erkenntnisse aus dem SPIEGEL studentenpisa-test.* Wiesbaden: VS Verlag.