UNIVERSITÀ DEGLI STUDI DI UDINE

DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

On Wald Tests for Differential Item Functioning Detection

Michela Battauz

February 2017

n. 3/2017

Statistica Metodologica e Sperimentale

On Wald Tests for Differential Item Functioning Detection

Michela Battauz
Department of Economics and Statistics, University of Udine

Abstract

Wald-type tests are a common procedure among the IRT-based methods for DIF detection. However, the empirical type I error rate of these tests departs from the significance level. In this paper, two reasons that explain this discrepancy will be discussed and a new procedure will be proposed. The first reason is related to the equating coefficients used to convert the item parameters to a common scale, as they are treated as known constants whereas they are estimated. The second reason is related to the parameterization used to estimate the item parameters, which is different from the usual IRT parameterization. Since the item parameters in the usual IRT parameterization are obtained in a second step, the corresponding covariance matrix is approximated using the delta method. The proposal of this article is to account for the estimation of the equating coefficients treating them as random variables and to use the untransformed (i.e. not reparameterized) item parameters in the computation of the test statistics. A simulation study is presented to compare the performance of this new proposal with the currently used procedure. Results show that the new proposal gives type I error rates closer to the significance level.

1 Introduction

Differential Item Functioning (DIF) is a violation of the invariance assumption in Item Response Theory (IRT) models and occurs when the probability of a positive response for examinees at the same ability level varies in different groups. Various methods have been proposed in the literature for the detection of DIF (see for example Magis et al., 2010). Among them, the Lord's chi-square test (Lord, 1980) is a common procedure that presents the advantage of requiring the estimation of the item parameters just ones for each group, as the selection of the anchor items is performed in a second step. The test was originally developed for detecting DIF between two groups, and then extended to the case of multiple groups by Kim et al. (1995). However, simulation studies reported in the literature showed that the empirical type I error rates for this test departs from the significance level (Kim et al., 1994). In particular, they are largely greater than the significance level for the Three-Parameter Logistic (3PL) model, while they are smaller for the Two-Parameter Logistic (2PL) model. In this paper, the reasons of this

discrepancy will be discussed, and a new proposal will be presented. The new proposal applies to multiple groups as well as to two groups.

Two alternative procedures can be found in Woods et al. (2013). The Wald-1 procedure relies on simultaneous estimation of item parameters, constraining the anchor items to have equal values across the groups. Instead, in the Wald-2 procedure the item parameters are estimated separately in different groups, fixing the mean and standard deviation of the abilities in the focus groups to the values previously estimated from concurrent calibration. While the type I error is extremely inflated for Wald-2, Wald-1 performs very well in simulations. However, this procedure requires the selection of the anchor items a priori, and the results may be affected by the presence of DIF in anchor items (Candell and Drasgow, 1988). Instead, the procedure presented in this paper keeps separate the estimation of item parameters from the computation of the test statistics, thus allowing to select the set of anchor items in a second iterative step, on the basis of the classification of the items as DIF.

This paper is structured as follows. Section 2 reviews the traditional Lord's Chi-Square test and its extension to multiple groups. Section 3 illustrates the new proposal, which is compared to the traditional procedure by means of simulation studies in Section 4. Finally, Section 5 contains some concluding remarks.

2 Review of Wald tests for DIF

In a 3PL model, the probability of a correct response to item j for a subject with ability θ is given by

$$p_j(\theta; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\{Da_j(\theta - b_j)\}}{1 + \exp\{Da_j(\theta - b_j)\}}, \tag{1}$$

where a_j, b_j and c_j are the discrimination, difficulty and guessing parameters. The 2PL model is obtained when the guessing parameters c_j are set to zero, while the Rasch model requires also that the discrimination parameters are equal to 1. The item parameters are generally estimated by means of the marginal maximum likelihood method (Bock and Aitkin, 1981).

Let $v_{jk} = (a_{jk}, b_{jk}, c_{jk})^\top$ be the vector of item parameters for group k. The Lord's Chi-square test was originally formulated for the case of two groups under investigation. The null hypothesis is the invariance of item parameters across groups

$$H_0 : \begin{pmatrix} a_{j1} \\ b_{j1} \\ c_{j1} \end{pmatrix} = \begin{pmatrix} a_{j2} \\ b_{j2} \\ c_{j2} \end{pmatrix}.$$

Without loss of generality, throughout this paper it is assumed that the reference group is group 1.

When the item parameters are estimated separately for the two groups, item parameter estimates are expressed on different measurement scales due to identifiability issues. Before comparing item parameter estimates deriving from different groups, it is then necessary to transform them in order to obtain values expressed on the same

metric. The equating transformations that permit to transform the item parameters estimates from the scale of group k to the scale of the reference group are

$$\hat{a}^*_{jk} = \frac{\hat{a}_{jk}}{A_k}, \tag{2}$$

and

$$\hat{b}^*_{jk} = A_k\,\hat{b}_{jk} + B_k, \tag{3}$$

where A_k and B_k are two constants called equating coefficients (Kolen and Brennan, 2014). The guessing parameters c_j do not need to be transformed. The test statistics is

$$\chi^2_j = (v_{j1} - v^*_{j2})^\top (\Sigma_{j1} + \Sigma^*_{j2})^{-1} (v_{j1} - v^*_{j2}), \tag{4}$$

where the vector of estimates of the parameters of item j in group k is

$$v_{jk} = (\hat{a}_{jk}, \hat{b}_{jk}, \hat{c}_{jk})^\top,$$

the vector of estimates transformed to the scale of the reference group is

$$v^*_{jk} = (\hat{a}^*_{jk}, \hat{b}^*_{jk}, \hat{c}_{jk})^\top,$$

Σ_{jk} is the estimated covariance matrix of v_{jk} and Σ^*_{jk} is the estimated covariance matrix of v^*_{jk}.

Kim et al. (1995) extended the test to the case of multiple groups, considering as null hypothesis

$$H_0 : \begin{pmatrix} a_{j1} \\ b_{j1} \\ c_{j1} \end{pmatrix} = \cdots = \begin{pmatrix} a_{jk} \\ b_{jk} \\ c_{jk} \end{pmatrix} = \cdots = \begin{pmatrix} a_{jK} \\ b_{jK} \\ c_{jK} \end{pmatrix} \tag{5}$$

and as test statistics

$$Q_j = (Cv_j)^\top (C\Sigma_j C^\top)^{-1} (Cv_j), \tag{6}$$

where

$$v_j = (v_{j1}^\top, {v^*_{j2}}^\top, \ldots, {v^*_{jK}}^\top)^\top,$$
$$\Sigma_j = \mathsf{COV}(v_j) = \text{blockdiag}(\Sigma_{j1}, \Sigma^*_{j2}, \ldots, \Sigma^*_{jK}),$$

blockdiag$(\cdot)$ denotes a block diagonal matrix and C is a contrast matrix. When $K = 2$, Equation (6) returns the test statistics (4). Under the null hypothesis, the asymptotic distribution of the test statistics is a Chi-square distribution with degrees of freedom equal to the number of rows of the matrix C.

3 A new proposal

Simulation studies reported in the literature (Kim et al., 1994) showed that the empirical type I error rate for this test diverges from the significance level. In particular, it is largely greater for the 3PL model, while it is smaller for the 2PL model.

The proposal of this paper aims at narrowing the discrepancy between the empirical type I error rate and the nominal value. Two issues will be considered to this end.

First, the equating coefficients in Equation (2) and (3) are treated as known constants in the computation of Σ^*_{jk}, while they are actually estimated (see for example Kim et al., 1994, 1995). The literature on test equating provides various methods for the estimation of the equating coefficients (Kolen and Brennan, 2014), while the asymptotic standard errors are derived in Ogasawara (2000) and Ogasawara (2001). The proposal of this paper is to account for the estimation of the equating coefficients in the computation of the covariance matrix of the item parameters.

A second issue regards the parameterization usually used for the estimation of the item parameters, which is

$$p_j(\theta;\gamma_j,\beta_{1j},\beta_{2j}) = c_j + (1-c_j)\frac{\exp(\beta_{1j}+\beta_{2j}\theta)}{1+\exp(\beta_{1j}+\beta_{2j}\theta)}, \tag{7}$$

with

$$c_j = \frac{\exp(\gamma_j)}{1+\exp(\gamma_j)}. \tag{8}$$

The set of parameters estimated for each item is then $\{\gamma_j,\beta_{1j},\beta_{2j}\}$, while the parameters of the usual IRT parameterization given in (1) are obtained using these transformations:

$$a_j = \frac{\beta_{2j}}{D} \tag{9}$$

$$b_j = -\frac{\beta_{1j}}{\beta_{2j}} \tag{10}$$

and Equation (8). The covariance matrices Σ_{jk} are obtained by applying the delta method.

Of course, the item parameter estimates need to be converted to a common metric. This can be performed using the following equations:

$$\hat{\beta}^*_{2jk} = \frac{\hat{\beta}_{2jk}}{\hat{A}_k}, \tag{11}$$

and

$$\hat{\beta}^*_{1jk} = \hat{\beta}_{1jk} - \hat{\beta}_{2jk}\frac{\hat{B}_k}{\hat{A}_k}. \tag{12}$$

The derivation is given in Appendix A.

The proposal of this paper is to compute the test statistics using untransformed item parameter estimates. The null hypothesis is then

$$H_0 : \begin{pmatrix}\gamma_{j1}\\ \beta_{1j1}\\ \beta_{2j1}\end{pmatrix} = \cdots = \begin{pmatrix}\gamma_{jk}\\ \beta_{1jk}\\ \beta_{2jk}\end{pmatrix} = \cdots = \begin{pmatrix}\gamma_{jK}\\ \beta_{1jK}\\ \beta_{2jK}\end{pmatrix}, \tag{13}$$

which is equivalent to (5). The test statistics is given by

$$W_j = (C\nu_j)^\top (C\Omega_j C^\top)^{-1}(C\nu_j), \tag{14}$$

where

$$\nu_j = (\nu_{j1}^\top, {\nu_{j2}^*}^\top, \ldots, {\nu_{jK}^*}^\top)^\top,$$

$$\nu_{jk} = (\hat{\gamma}_{jk}, \hat{\beta}_{1jk}, \hat{\beta}_{2jk})^\top, \quad \nu_{jk}^* = (\hat{\gamma}_{jk}, \hat{\beta}_{1jk}^*, \hat{\beta}_{2jk}^*)^\top,$$

$$\Omega_j = \mathsf{COV}(\nu_j)$$

and C is a contrast matrix. It is important to note that, accounting for the estimation of the equating coefficients, not only the covariance matrix of ν_{jk}^* needs to be property calculated, but also the covariance matrices between ν_{j1} and ν_{jk}^* are not zero. This is because the equating coefficients are estimated using the item parameter estimates obtained from group 1 and group k. For more details on the computation of the covariance matrix Ω_j, see Appendix B.

4 Simulation studies

The performance of the new proposal was assessed by means of simulation studies. Various settings were considered. The IRT models used to generate the data and estimate the item parameters are the 2PL and the 3PL models. The sample size for each group takes values $n = \{500, 1000, 2000, 4000\}$, while the number of items of the test is 20 and 40. Test responses of 3 groups were simulated. For each group the θ values were generated from a normal distribution with mean $\{0, 0.5, -0.5\}$ and standard deviation $\{1, 1.2, 0.8\}$ in the 3 groups. The discrimination parameters were generated from a uniform distribution with range $[0.7, 1.3]$; the difficulty parameters were generated from a standard normal distribution and the guessing parameters were taken equal to 0.2. The percentage of DIF items was 0%, 5% and 20%. In presence of DIF, the values added to the item parameters in the two focus groups were 0.3 and 0.5 for the discrimination parameters, and 0.4 and 0.6 for the difficulty parameters. For each setting, 1000 simulated data sets were generated. The test was applied to 2 and 3 groups (the third group was excluded when just 2 groups were considered). The traditional test was also performed for comparison. The purification procedure (Candell and Drasgow, 1988) was applied in presence of DIF items. The new and the traditional procedures were implemented in R (R Development Core Team, 2016), employing the `equateIRT` package (Battauz, 2015) for the computation of the equating coefficients (see the program code in the Supplementary Material supplied with the online version of this paper). The R package `ltm` was used to estimate the IRT models (Rizopoulos, 2006).

The data sets simulated without DIF items are used to evaluate the type I error rates. The empirical type I error rates are reported in Table 1, while Figures 1 and 2 give a graphical representation for the 2PL and the 3PL models respectively. Consistently with previous studies, using the traditional procedure, the type I error rate is lower than the significance level for the 2PL model and larger for the 3PL model. The new procedure provides instead values much closer to the nominal level under all the settings considered. There are only a couple of exceptions for the 3PL model with 40 items and 2 groups, where the traditional procedure performed better. However, under this setting, the new procedure still performs better for smaller sample sizes and the departure from the nominal level is anyway small for larger sample sizes.

Figure 1: Type I error rates (false positive rate) for the 2PL model.

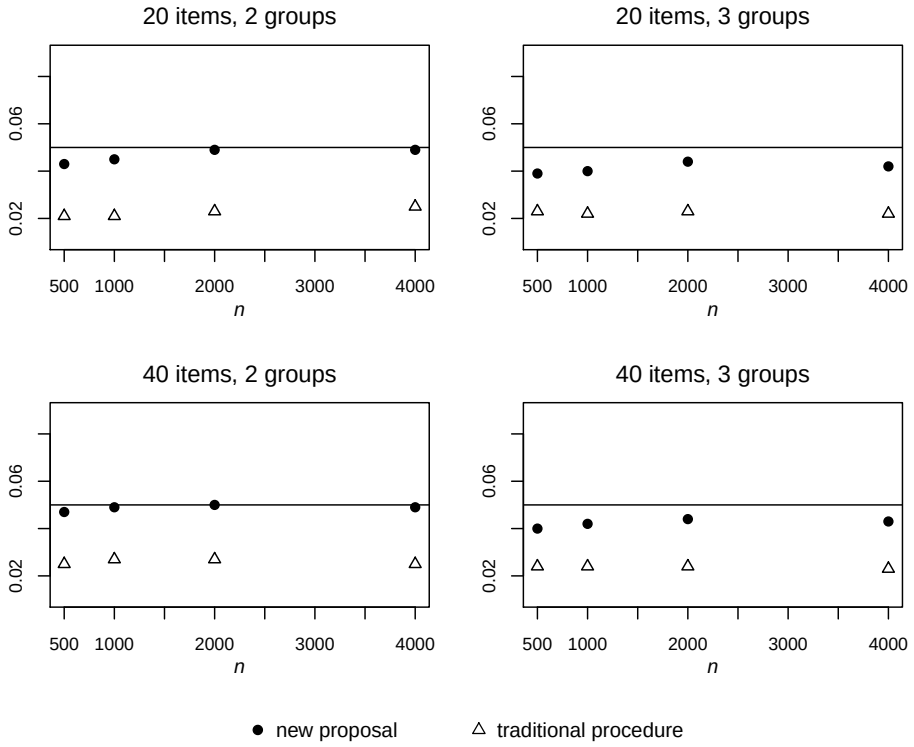


Figure 2: Type I error rates (false positive rate) for the 3PL model.

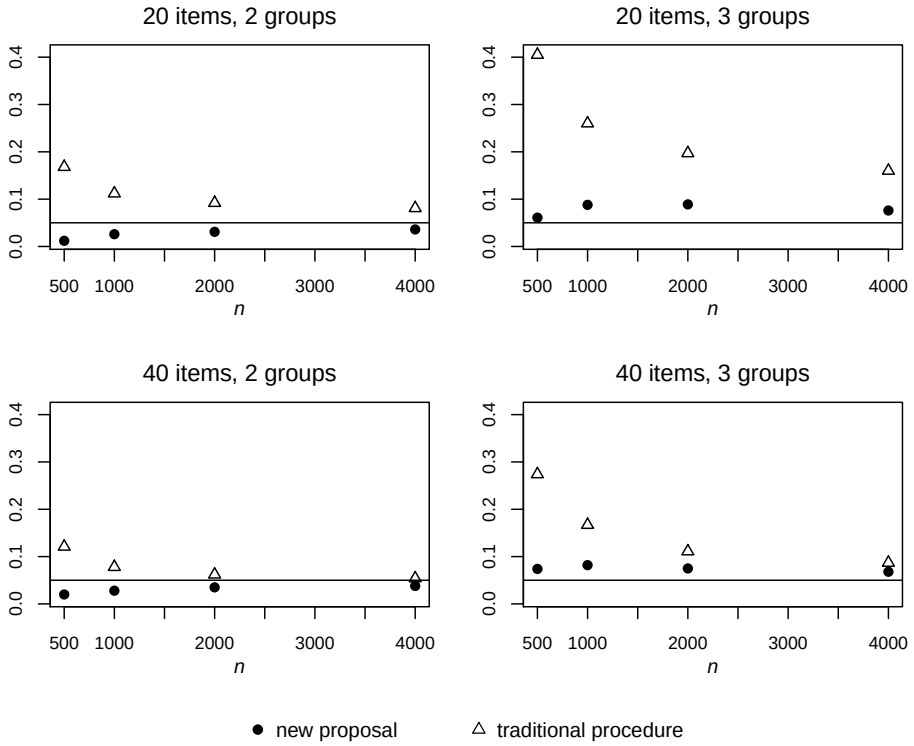


Table 1: Type I error rate (false positive rate).

			2 groups		3 groups	
model	test length	n	new	traditional	new	traditional
2PL	20	500	0.043	0.021	0.039	0.023
2PL	20	1000	0.045	0.021	0.040	0.022
2PL	20	2000	0.049	0.023	0.044	0.023
2PL	20	4000	0.049	0.025	0.042	0.022
2PL	40	500	0.047	0.025	0.040	0.024
2PL	40	1000	0.049	0.027	0.042	0.024
2PL	40	2000	0.050	0.027	0.044	0.024
2PL	40	4000	0.049	0.025	0.043	0.023
3PL	20	500	0.012	0.168	0.061	0.405
3PL	20	1000	0.026	0.112	0.088	0.260
3PL	20	2000	0.031	0.092	0.089	0.197
3PL	20	4000	0.036	0.081	0.076	0.160
3PL	40	500	0.020	0.121	0.074	0.274
3PL	40	1000	0.028	0.078	0.082	0.167
3PL	40	2000	0.035	0.062	0.075	0.111
3PL	40	4000	0.038	0.054	0.068	0.087

When test responses are simulated in presence of DIF, it is also possible to evaluate the power of the test. Figures 3 and 4 represent the empirical power of the tests with a percentage of 5% of DIF items. Figure 3 shows that for the 2PL model the power is substantially equal to 1 under all the settings for both the procedures. For the 3PL model (Figure 4), the power of the traditional procedure is higher for smaller sample sizes, and it tends to 1 as the sample size increases for both the procedures. However, it should be noted that a comparison is not appropriate since the type I error rates are not equal for the two procedures, and a greater power should be expected from a test that tends to reject the null hypothesis too often.

Results for the case of a percentage of DIF items equal to 20% are not shown because very similar to the case of a percentage of 5%.

Figure 3: Power (true positive rate) for the 2PL model (percentage of DIF items: 5%.).

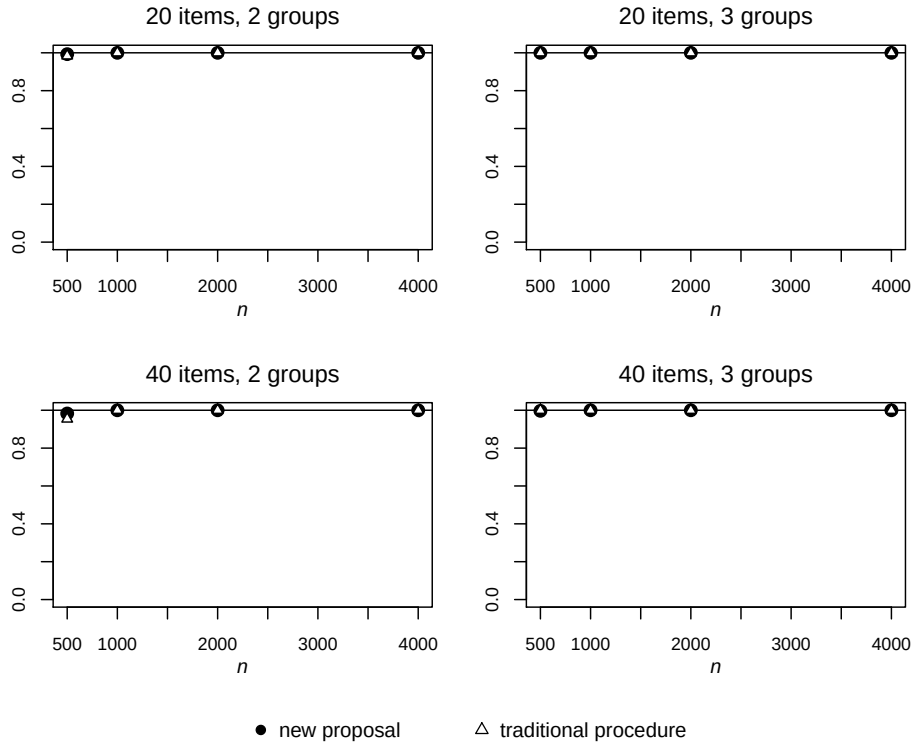
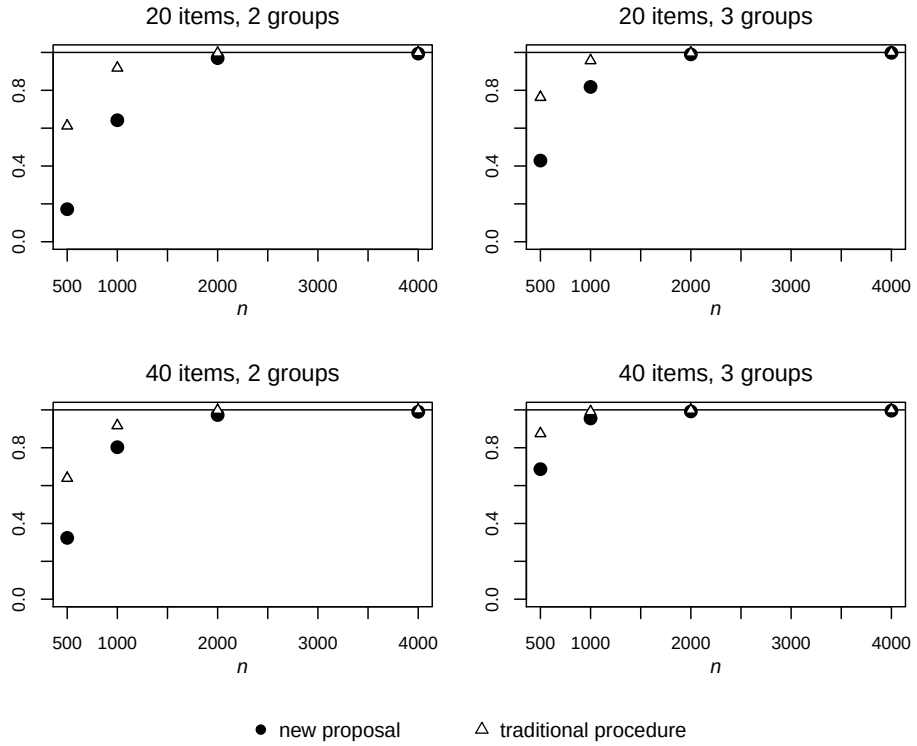


Figure 4: Power (true positive rate) for the 3PL model (percentage of DIF items: 5%.).



5 Discussion

In this article a new procedure to perform Wald-type tests for DIF detection is presented. The new procedure recognizes two basic aspects. One is the random nature of the estimated equating coefficients, which should be taken into account for an accurate computation of the covariance matrix. Another issue is the non-invariance to a non-linear reparametrization of the Wald test, (Gregory and Veall, 1985). Thus, not applying any unnecessary reparameterization to the item parameters estimated with the marginal maximum likelihood method is certainly preferable. The simulation studies presented in this paper showed that the new proposal outperforms the traditional procedure. The results are better for the 2PL model than the 3PL model, and a sensible explanation for this difference can found in the difficulties of maximum likelihood fitting algorithms for the 3PL model (Patz and Junker, 1999).

Appendix A: Equating of untransformed item parameters

Equation (11) is obtained from Equations (2) and (9) as follows:

$$\hat{\beta}^*_{2jk} = D\hat{a}^*_{jk} = \frac{D\hat{a}_{jk}}{\hat{A}_k} = \frac{\hat{\beta}_{2jk}}{\hat{A}_k}. \tag{A1}$$

Equations (2), (3) and (10) lead to Equation (12):

$$\hat{\beta}^*_{1jk} = -D\hat{a}^*_{jk}\hat{b}^*_{jk} = -D\frac{\hat{a}_{jk}}{\hat{A}_k}(\hat{A}_k\,\hat{b}_{jk} + \hat{B}_k) = -D\hat{a}_{jk}\hat{b}_{jk} - D\hat{a}_{jk}\frac{\hat{B}_k}{\hat{A}_k} = \hat{\beta}_{1jk} - \hat{\beta}_{2jk}\frac{\hat{B}_k}{\hat{A}_k}. \tag{A2}$$

Appendix B: Covariance matrix of item parameters

The covariance matrix Ω_j entering in Equation (14) is a block matrix given by

$$\Omega_j = \begin{pmatrix} \mathsf{COV}(\nu_{j1}) & \mathsf{COV}(\nu_{j1}, \nu^*_{j2}) & \mathsf{COV}(\nu_{j1}, \nu^*_{j3}) & \dots & \mathsf{COV}(\nu_{j1}, \nu^*_{jK}) \\ \mathsf{COV}(\nu^*_{j2}, \nu_{j1}) & \mathsf{COV}(\nu^*_{j2}) & 0 & \dots & 0 \\ \mathsf{COV}(\nu^*_{j3}, \nu_{j1}) & 0 & \mathsf{COV}(\nu^*_{j3}) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathsf{COV}(\nu^*_{jK}, \nu_{j1}) & 0 & 0 & \dots & \mathsf{COV}(\nu^*_{jK}) \end{pmatrix}.$$

Let Ω_{jk} denote $\mathsf{COV}(\nu_{jk})$, which is obtained from the estimation of the item parameters. Using the delta method, it is possible to find the covariance matrix

$$\begin{aligned}
\mathsf{COV}((\nu_{j1}^\top, \nu_{jk}^{*\top})^\top) &= \frac{\partial(\nu_{j1}^\top, \nu_{jk}^{*\top})^\top}{\partial(\nu_{j1}^\top, \nu_{jk}^\top)} \mathsf{COV}((\nu_{j1}^\top, \nu_{jk}^\top)^\top) \frac{\partial(\nu_{j1}^\top, \nu_{jk}^{*\top})}{\partial(\nu_{j1}^\top, \nu_{jk}^\top)^\top} \\
&= \begin{pmatrix} \frac{\partial \nu_{j1}}{\partial \nu_{j1}^\top} & \frac{\partial \nu_{j1}}{\partial \nu_{jk}^\top} \\ \frac{\partial \nu_{jk}^*}{\partial \nu_{j1}^\top} & \frac{\partial \nu_{jk}^*}{\partial \nu_{jk}^\top} \end{pmatrix} \begin{pmatrix} \Omega_{j1} & 0 \\ 0 & \Omega_{j2} \end{pmatrix} \begin{pmatrix} \frac{\partial \nu_{j1}^\top}{\partial \nu_{j1}} & \frac{\partial \nu_{jk}^{*\top}}{\partial \nu_{j1}} \\ \frac{\partial \nu_{j1}^\top}{\partial \nu_{jk}} & \frac{\partial \nu_{jk}^{*\top}}{\partial \nu_{jk}} \end{pmatrix} \\
&= \begin{pmatrix} \Omega_{j1} & \Omega_{j1} \frac{\partial \nu_{jk}^{*\top}}{\partial \nu_{j1}} \\ \frac{\partial \nu_{jk}^*}{\partial \nu_{j1}^\top} \Omega_{j1} & \frac{\partial \nu_{jk}^*}{\partial \nu_{j1}^\top} \Omega_{j1} \frac{\partial \nu_{jk}^{*\top}}{\partial \nu_{j1}} + \frac{\partial \nu_{jk}^*}{\partial \nu_{jk}^\top} \Omega_{j2} \frac{\partial \nu_{jk}^{*\top}}{\partial \nu_{jk}}, \end{pmatrix}
\end{aligned}$$

since $\frac{\partial \nu_{j1}}{\partial \nu_{j1}^\top}$ is the identity matrix and $\frac{\partial \nu_{j1}}{\partial \nu_{jk}^\top} = 0$. The blocks on the main diagonal of Ω_j are then

$$\mathsf{COV}(\nu_{jk}^*) = \frac{\partial \nu_{jk}^*}{\partial \nu_{j1}^\top} \Omega_{j1} \frac{\partial \nu_{jk}^{*\top}}{\partial \nu_{j1}} + \frac{\partial \nu_{jk}^*}{\partial \nu_{jk}^\top} \Omega_{j2} \frac{\partial \nu_{jk}^{*\top}}{\partial \nu_{jk}},$$

while the non-zero matrices outside the main diagonal are given by

$$\mathsf{COV}(\nu_{j1}, \nu_{jk}^*) = \Omega_{j1} \frac{\partial \nu_{jk}^{*\top}}{\partial \nu_{j1}}.$$

The chain rule can be exploited to find the derivatives

$$\frac{\partial \nu_{jk}^*}{\partial(\nu_{j1}^\top, \nu_{jk}^\top)} = \frac{\partial \nu_{jk}^*}{\partial(\nu_{jk}^\top, \hat{A}_k, \hat{B}_k)} \frac{\partial(\nu_{jk}^\top, \hat{A}_k, \hat{B}_k)^\top}{\partial(\nu_{j1}^\top, \nu_{jk}^\top)}, \tag{B1}$$

where

$$\frac{\partial(\hat{A}_k, \hat{B}_k)^\top}{\partial(\nu_{j1}^\top, \nu_{jk}^\top)} = \frac{\partial(\hat{A}_k, \hat{B}_k)^\top}{\partial(v_{j1}^\top, v_{jk}^\top)} \frac{\partial(v_{j1}^\top, v_{jk}^\top)^\top}{\partial(\nu_{j1}^\top, \nu_{jk}^\top)}. \tag{B2}$$

The non-zero derivatives entering in (B1) and (B2) are given in the following (derivatives of a variable with respect to itself are not shown):

$$\begin{aligned}
&\frac{\partial \hat{\beta}_{1jk}^*}{\partial \hat{\beta}_{1jk}} = 1, \quad \frac{\partial \hat{\beta}_{1jk}^*}{\partial \hat{\beta}_{2jk}} = -\frac{\hat{B}_k}{\hat{A}_k}, \quad \frac{\partial \hat{\beta}_{1jk}^*}{\partial \hat{A}_k} = \hat{\beta}_{2jk} \frac{\hat{B}_k}{\hat{A}_k^2}, \\
&\frac{\partial \hat{\beta}_{1jk}^*}{\partial \hat{B}_k} = \frac{\hat{\beta}_{2jk}}{\hat{A}_k}, \quad \frac{\partial \hat{\beta}_{2jk}^*}{\partial \hat{\beta}_{2jk}} = \frac{1}{\hat{A}_k}, \quad \frac{\partial \hat{\beta}_{2jk}^*}{\partial \hat{A}_k} = -\frac{\hat{\beta}_{2jk}}{\hat{A}_k^2} \\
&\frac{\partial \hat{a}_{jk}}{\partial \hat{\beta}_{2jk}} = \frac{1}{D}, \quad \frac{\partial \hat{b}_{jk}}{\partial \hat{\beta}_{1jk}} = -\frac{1}{\hat{\beta}_{2j1}}, \quad \frac{\partial \hat{b}_{jk}}{\partial \hat{\beta}_{2jk}} = \frac{\hat{\beta}_{1jk}}{\hat{\beta}_{2jk}^2}.
\end{aligned}$$

The derivatives $\frac{\partial(\hat{A}_k, \hat{B}_k)^\top}{\partial(v_{j1}^\top, v_{j2}^\top)}$ are given in Ogasawara (2000) and Ogasawara (2001).

References

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, 68(7):1–22. doi=10.18637/jss.v068.i07.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459. doi: 10.1007/BF02293801.

Candell, G. L. and Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied psychological measurement*, 12(3):253–260. doi: 10.1177/014662168801200304.

Gregory, A. W. and Veall, M. R. (1985). Formulating Wald tests of nonlinear restrictions. *Econometrica*, 53(6):1465–1468. doi: 10.2307/1913221.

Kim, S.-H., Cohen, A. S., and Kim, H.-O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement*, 18(3):217–228. doi: 10.1177/014662169401800303.

Kim, S.-H., Cohen, A. S., and Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32(3):261–276. doi: 10.1111/j.1745-3984.1995.tb00466.x.

Kolen, M. and Brennan, R. (2014). *Test Equating, Scaling, and Linking: Methods and Practices (3rd ed.)*. New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Magis, D., Béland, S., Tuerlinckx, F., and De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42(3):847–862. doi:10.3758/BRM.42.3.847.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce)*, 51(1):1–23.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, 25(1):53–67. doi: 10.1177/01466216010251004.

Patz, R. J. and Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of educational and behavioral statistics*, 24(4):342–366. doi: 10.3102/10769986024004342.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software*, 17(5):1–25. doi:10.18637/jss.v017.i05.

Woods, C. M., Cai, L., and Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, 73(3):532–547. doi: 10.1177/0013164412464875.