Detection of Differential Item Functioning via the Credible Intervals and Odds Ratios Methods

Ya-Hui Su and Henghsiu Tsai

Abstract Differential item functioning (DIF) analysis is an essential procedure for educational and psychological tests to identify items that exhibit varying degrees of DIF. DIF means that the assumption of measurement invariance is violated, and then test scores are incomparable for individuals of the same ability level from different groups, which substantially threatens test validity. In this paper, we investigated the credible intervals (CI) and odds ratios (OR) methods to detect uniform DIF within the framework of the Rasch model through a series of simulations. The results showed that the CI method performed better than the OR method to identify DIF items under the balanced DIF conditions. However, the CI method yielded inflated false positive rates under the unbalanced DIF conditions. The effectiveness of these two approaches was illustrated with an empirical example.

Keywords Credible interval · Odds ratio · DIF · Markov chain Monte Carlo · IRT

1 Introduction

Differential item functioning (DIF) analysis is an essential procedure for educational and psychological tests. DIF occurs when individuals from different groups (such as gender, ethnicity, country, or age) have different probabilities of endorsing or accurately answering a given item after controlling for overall test scores. It violates the assumption of measurement invariance and the test scores become incomparable for individuals of the same ability level from different groups, which substantially threatens test validity. DIF detection can examine how test scores are affected by

Y.-H. Su
Department of Psychology, National Chung Cheng University, 168 University Road, Minhsiung Township, 62102 Chiayi County, Taiwan
e-mail: psyyhs@ccu.edu.tw

H. Tsai (✉)
Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang District, 11529 Taipei, Taiwan
e-mail: htsai@stat.sinica.edu.tw

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_28

external variables that are not related to the construct (Glas, 1998). Therefore, it is important to know if items are subject to DIF; that is, to know if the examinees are fairly measured.

Many approaches have been developed to perform DIF detection, and they can be classified into two categories (Magis, Béland, Tuerlinckx, & De Boeck, 2010): item response theory (IRT)-based and non-IRT-based approaches. The IRT-based approaches include the Lagrange multiplier test (Glas, 1998), the likelihood ratio test (Cohen, Kim, & Wollack, 1996), Lord's chi-square test (Lord, 1980), Raju's (1988) signed area method, etc. The IRT-based approaches require estimating item parameters for different groups. After comparing these item parameters of different groups, an item is identified as a DIF item if the item parameters are significantly different between groups. By contrast, the non-IRT-based approaches require neither specific forms for the IRT models nor large sample sizes (Narayanon & Swaminathan, 1996). The non-IRT-based approaches include the Mantel-Haenszel (MH; Holland & Thayer, 1988), logistic regression (LR; Rogers & Swaminathan, 1993), simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) methods, etc.

Among the non-IRT-based approaches, the MH and LR methods perform well in flagging DIF items when the percentage of DIF items is not very high and there is no mean ability difference between groups (French & Maller, 2007; Narayanon & Swaminathan, 1996). A common feature of these two methods is that examinees from different groups are placed on a common metric based on the test scores, which are usually called matching variables. The use of the matching variables is critical for DIF detection (Kopf, Zeileis, & Strobl, 2015). If the matching variables are contaminated (i.e., consisting of DIF items), examinees with the same ability levels would not be matched well, and the subsequent DIF detection would be biased (Clauser, Mazor, & Hambleton, 1993). In practice, it is challenging to identify a set of DIF-free items as the matching variables for DIF detection, especially when the percentage of DIF items is high or when DIF magnitudes are large (Narayanon & Swaminathan, 1996; Rogers & Swaminathan, 1993).

To overcome this difficulty, the odds ratios (OR; Jin, Chen, & Wang, 2018) method was proposed to detect uniform DIF under various manipulated variables, such as different DIF pattern, impact, sample size, and with/without purification. Jin, Chen, and Wang (2018) found that the OR method without a purification procedure outperformed the MH and LR methods in controlling false positive rates (FPR) and obtaining high true positive rates (TPR) when tests contained high percentages of DIF items. Another recently developed IRT-based DIF detection method was the credible interval (CI) method proposed by Su, Chang, & Tsai (2018) to detect uniform and non-uniform DIF items under the Bayesian framework. Su et al (2018) found that the CI method performed well; however, only unbalanced DIF conditions and no impact (i.e., mean ability difference between the reference and focal groups was zero) were considered in their study.

A common feature of the CI and OR methods is that both methods perform DIF detection after constructing intervals. The OR method follows the frequentist approach, and constructs the confidence interval for the mean ability difference between the reference and focal groups. By contrast, the CI method follows the

Bayesian approach, and constructs the credible interval for the item difficulty difference between the reference and focal groups. See next section for more details. Because of the nature of the Bayesian framework, the CI method would need more time to perform DIF examination. Besides, the CI method assumes Rasch (1960) model is a correct model for the data. By contrast, the OR method does not require the specification of an IRT model; however, this method may not work when the number of examinees of any group is very small. Given the very different nature of these two newly developed methods, it is interesting to compare these two methods under the Rasch model. In this paper, we investigated the performance of the CI and OR methods to detect uniform DIF within the framework of the Rasch model through a series of simulation studies. The effectiveness of these two approaches was illustrated with an empirical example.

2 The CI and OR DIF Detection Methods

2.1 The CI Method

We first review the CI method proposed by Su, Chang, and Tsai (2018). Let Y_{pj} be the dichotomous response of examinee p on item j, where $p = 1, \ldots, P$, and $j = 1, \ldots, J$. Denote b_j and θ_p as the difficulty parameter for item j and the examinee ability parameter for examinee p, respectively. In the Rasch (1960) model, the probability of examinee p getting a correct response on item j is given by

$$\pi_{pj} = \mathrm{P}\big(Y_{pj} = 1 | \theta_p, b_j\big) = \frac{1}{1 + e^{-\theta_p + b_j}}. \tag{1}$$

An item is flagged as DIF if the probability of answering the item correctly differs across different groups after controlling for the underlying ability levels. The CI method was proposed to perform DIF detection under a Bayesian estimation framework (Su et al., 2018). Consider the simplest case of two groups, hence, examinee p either belongs to the reference group ($g_p = 0$) or to the focal group ($g_p = 1$). Furthermore, each group has its own difficulty parameter. Then, Eq. (1) becomes

$$\pi_{pj} = \mathrm{P}\big(Y_{pj} = 1 \big| g_p, \theta_p, b_j, d_j\big) = \begin{cases} \frac{1}{1+e^{-\theta_p + b_j}}, & g_p = 0, \\ \frac{1}{1+e^{-\theta_p + d_j}}, & g_p = 1, \end{cases} \tag{2}$$

where b_j and d_j are the difficulty parameters for the reference and the focal groups, respectively. Alternatively, the notations of Glas (1998) is adopted to rewrite Eq. (2) as

$$\pi_{pj} = \mathrm{P}\big(Y_{pj} = 1 \big| g_p, \theta_p, b_j, \delta_j\big) = \begin{cases} \frac{1}{1+e^{-\theta_p + b_j}}, & g_p = 0, \\ \frac{1}{1+e^{-\theta_p + b_j + \delta_j}}, & g_p = 1. \end{cases} \tag{3}$$

Equation (3) implies that the responses of the focal group need an additional difficulty parameter δ_j. Therefore, the following hypothesis is considered:

$$H_0 : \delta_j = 0 \text{ versus } H_1 : \delta_j \neq 0.$$

Due to the complexity of the likelihood function, a Bayesian estimation method is used. Specifically, we follow closely the Bayesian approaches proposed by Chang, Tsai, and Hsu (2014), Chang, Tsai, Su, and Lin (2016), and Su et al. (2018). In particular, a two-layer hierarchical prior is assumed for the model parameters to reduce the impact of the prior settings on the posterior inference. For model identification, we follow Frederickx, Tuerlinckx, de Boeck, and Magis (2010)'s paper by assuming that the marginal distribution of θ_p is normal:

$$\theta_p \sim \begin{cases} N\left(0, \sigma_r^2\right), & g_p = 0, \\ N\left(\mu_f, \sigma_f^2\right), & g_p = 1. \end{cases}$$

For the first-layer prior settings for the parameters, we assume

$$\begin{aligned} b_j &\sim N\left(\mu_b, \sigma_b^2\right), \\ d_j &\sim N\left(\mu_d, \sigma_d^2\right). \end{aligned}$$

Given the first-layer prior, we assume the second-layer prior to be

$$\begin{aligned} \mu_f &\sim N\left(\mu_1, \sigma_1^2\right), \\ \mu_b &\sim N\left(\mu_2, \sigma_2^2\right), \\ \mu_d &\sim N\left(\mu_3, \sigma_3^2\right), \\ \sigma_r^2 &\sim \text{Inv-Gamma}(\alpha_1, \beta_1), \\ \sigma_f^2 &\sim \text{Inv-Gamma}(\alpha_2, \beta_2), \\ \sigma_b^2 &\sim \text{Inv-Gamma}(\alpha_3, \beta_3), \\ \sigma_d^2 &\sim \text{Inv-Gamma}(\alpha_4, \beta_4). \end{aligned}$$

All parameters in the second-layer priors,

$$(\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4),$$

are assigned in a reasonable way. Furthermore, we also assume that all the priors are independent.

More specifically, the CI method proceeds as follows. There are J items in the test, and each of the J items in the test is examined one at a time. For item j, a size α test of $\delta_j = 0$ is constructed. Let item j follow Eq. (3) and the other items follow Eq. (1). That is, item j is tested if the responses of the focal group need an additional parameter δ_j. The Bayesian analysis via the Markov chain Monte Carlo (MCMC) scheme is implemented to construct the equal-tailed $1 - \alpha$ credible interval for the

parameter δ_j. If the interval includes 0, then $\delta_j = 0$ is not rejected. Otherwise, $\delta_j = 0$ is rejected, and hence item j is considered a DIF item.

2.2 The OR Method

The OR method was proposed by Jin, Chen, and Wang (2018) to detect uniform DIF. Let n_{R1j} and n_{R0j} be the numbers of examinees for the reference group who answer item j correctly and incorrectly, respectively; and let n_{F1j} and n_{F0j} be the numbers of examinees for the focal group who answer item j correctly and incorrectly, respectively. For item j, let $\hat{\lambda}_j$ denote the logarithm of the OR of success over failure for the reference and focal groups:

$$\hat{\lambda}_j = \log\left(\frac{n_{R1j}/n_{R0j}}{n_{F1j}/n_{F0j}}\right), \tag{4}$$

which follows a normal distribution asymptotically (Agresti, 2002) with mean λ and standard deviation

$$\sigma(\hat{\lambda}_j) = \left(n_{R1j}^{-1} + n_{R0j}^{-1} + n_{F1j}^{-1} + n_{F0j}^{-1}\right)^{1/2}, \tag{5}$$

where λ is the mean ability difference between the reference and focal groups. For each item j, $\hat{\lambda}_j$, $\sigma(\hat{\lambda}_j)$, and $\hat{\lambda}_j \pm z_{\alpha/2} \times \sigma(\hat{\lambda}_j)$ are computed. Then, find the median for $\hat{\lambda}_1, \hat{\lambda}_2, \ldots,$ and $\hat{\lambda}_J$. An item j is flagged as a DIF item if $\hat{\lambda}_j \pm z_{\alpha/2} \times \sigma(\hat{\lambda}_j)$, the $1-\alpha$ confidence interval of item j, does not cover the median of $\hat{\lambda}_1, \hat{\lambda}_2, \ldots,$ and $\hat{\lambda}_J$. Note that this method may not work when the number of examinees are very small because the values of $\hat{\lambda}_j$ cannot be computed when any numbers in Eq. (4) is zero. The scale purification procedures can easily be implemented with the OR method; all that is necessary is the precomputation of the sample median based on presumably DIF-free items. See Jin, Chen, and Wang (2018) for the details.

3 Simulation Study

3.1 Design

In this section, the simulation studies were conducted to compare the performance of the CI and OR methods. In each experiment, we simulated a test consisting of 20 items (i.e., $J = 20$). The number of examinee (P) is 1000. Specifically, we were interested in the comparisons based on the five factors, which were also considered in Simulation Study I of Jin et al. (2018). They were (a) equal and unequal sample sizes

of the reference and focal groups (500/500 and 800/200), (b) percentages of DIF items (0, 10, 20, 30 and 40%), (c) DIF patterns: balanced and unbalanced, (d) impact (0 and 1), and (e) purification procedure (with or without). Under the balanced DIF conditions, some DIF items favored the reference group and the other items favored the focal group. By contrast, under the unbalanced DIF conditions, all DIF items favored the reference group.

Item responses were generated according to Eq. (3). The true values of difficulty parameters b_j were generated identically and independently from a uniform distribution between -1.5 and 1.5. The true values of examinee ability parameters θ_p for the reference group ($g_p = 0$) were generated from the standard normal distribution. When impact $= 0$, the true values of θ_p for the focal group ($g_p = 1$) were also generated from the standard normal distribution; when impact $= 1$, they were generated from the normal distribution with mean -1 and variance 1. Under the unbalanced DIF conditions, $d_j - b_j = 0.5$ for all DIF items; under the balanced DIF conditions, $d_j - b_j = 0.5$ for the first half of the DIF items and $d_j - b_j = -0.5$ for the second half of the DIF items. We fixed α, the Type-I error of each test, to 0.05.

To construct the credible intervals, we produced 11,000 MCMC draws with the first 1000 draws as burn-in. A total of 100 replications were carried out under each condition. The performance of these two methods was compared in terms of the FPR and TPR. The FPR was the rate that DIF-free items were misclassified as having DIF whereas the TPR was rate that DIF items were correctly classified as having DIF. The averaged FPR across the DIF-free items and averaged TPR across the DIF items for these two methods were reported. Both the OR and CI methods were implemented by using FORTRAN code with IMSL subroutines, and are available upon request.

3.2 Results

The averaged FPR and TPR of two DIF detection methods for equal (500/500) and unequal (800/200) sample sizes list in Tables 1 and 2, respectively. As expected, both methods yielded well-controlled FPR under the no-DIF (0% DIF items) and balanced DIF conditions, although the OR method was slightly conservative. Similar to Jin, Chen, and Wang (2018)'s study, the FPR larger than or equal to 7.5% was defined as the inflated FPR in the present study. Under the unbalanced DIF conditions, the OR method yielded slightly inflated FPR only when tests had 40% or more DIF items. However, the CI method yielded inflated FPR when tests had 20% or more DIF items under the unbalanced DIF conditions. The TPR of the CI method was higher than that of the OR methods under two following conditions: (i) the balanced DIF conditions and (ii) the unbalanced DIF conditions with 10% DIF items. Furthermore, under these two conditions, the ratio of the TPR of the CI method to that of the OR method with scale purification procedure ranged from 1.01 to 1.27, and it was larger for unequal (800/200) sample sizes than that for equal (500/500) sample sizes. When the total sample size is 1000, the TPR for equal (500/500) sample sizes was higher than that for unequal (800/200) sample sizes. In general, both the FPR and TPR

increased with the percentages of DIF items. The TPR for the balanced DIF was higher than that for the unbalanced DIF, except for the OR method when Impact = 0 with equal (500/500) sample size. In general, the TPR was higher when Impact = 0 than that when Impact = 1. The purification procedure increased the TPR for the unbalanced DIF condition, and the higher the percentage of the DIF items, the higher the ratio of the TPR of the OR method with scale purification to that of the OR method without scale purification. By contrast, the purification procedure did not increase the TPR for the balanced DIF condition.

4 Application

In this section, the CI and OR methods described in the previous sections were applied to the data of the physics examination of the 2010 Department Required Test for college entrance in Taiwan provided by the College Entrance Examination Center (CEEC). Each examinee was required to answer 26 questions within 80 min. The 26 questions were further divided into three parts. The total score was 100, and the test was administered under the formula-scoring directions. For the first part, there were 20 multiple-choice questions, and the examinees had to choose one correct answer out of 5 possible choices. For each correct answer, 3 points were granted, and 3/4 point was deducted from the raw score for each incorrect answer. The second part consisted of 4 multiple-response questions, and each question consisted of 5 choices, examinees needed to select all the answer choices that apply. The choices in each item were knowledge-related, but were answered and graded separately. For each correct choice, 1 point was earned, and for each incorrect choice 1 point was deducted from the raw score. The final adjusted scores for each of these two parts started from 0. The last part consisted of 2 calculation problems, and deserved 20 points in total.

The data from 1000 randomly sampled examinees contained the original responses and nonresponses information, but we treated both nonresponses and incorrect answers the same way and coded them as $Y_{pj} = 0$ as Chang et al. (2014) suggested. As for the calculation part, the response Y_{pj} was coded as 1 whenever the original score was more than 7.5 out of 10 points, and zero otherwise (see also Chang et al., 2014). Here, we considered male and female as the reference and focal groups, respectively. Among the 1000 examinees, 692 of them were male and the others were female.

We made more MCMC draws than that in Sect. 3. Specifically, we produced 40,000 MCMC draws with the first 10,000 draws as burn-in. Then we tested $\delta_j = 0$, for $j = 1, \ldots, 26$. Again, we considered $\alpha = 0.05$. The intervals of $\hat{\lambda}_j \pm z_{\alpha/2} \times \sigma(\hat{\lambda}_j)$ for the OR method, which were the same for both with and without purification, and the credible intervals obtained from the real data were summarized in Table 3. Note that the median of $\hat{\lambda}_1, \hat{\lambda}_2, \ldots,$ and $\hat{\lambda}_J$ before and after purification were 0.5687 and 0.6163, respectively, so the OR method identified Items 3, 5, 8, 19 and 23 as DIF items, which were underlined and bolded in Table 3. Table 3 also showed that the

Table 1 Averaged FPR (%) and TPR (%) under the conditions with sample sizes of the reference and the focal groups: 500/500

			FPR			TPR		
Impact	DIF type	DIF (%)	OR with purification	OR without purification	CI	OR with purification	OR without purification	CI
Impact = 0		0	3.35	3.15	5.05	–	–	–
	Balanced DIF	10	3.66	3.22	5.06	85.00	85.50	93.50
		20	3.94	3.63	5.44	84.75	85.25	93.75
		30	3.71	3.43	5.43	86.67	86.67	92.17
		40	4.33	3.33	5.58	88.00	89.25	92.00
	Unbalanced DIF	10	4.06	3.67	5.83	87.00	87.00	88.00
		20	4.63	4.44	**10.56**	87.25	83.00	80.25
		30	4.71	5.50	**17.43**	84.33	76.33	70.67
		40	**7.83**	**10.58**	**29.00**	76.50	63.25	56.50
Impact=1		0	3.10	2.95	5.35	–	–	–
	Balanced DIF	10	2.89	2.83	5.39	81.00	81.00	89.50
		20	3.13	2.94	5.38	86.25	86.75	89.25
		30	3.36	3.00	5.29	84.17	85.50	89.33
		40	3.50	2.92	5.33	82.88	85.50	90.50
	Unbalanced DIF	10	3.28	3.11	6.28	78.00	77.00	80.50
		20	4.50	4.13	**10.38**	77.00	73.00	73.50
		30	5.07	6.64	**15.86**	72.67	65.50	65.33
		40	**9.17**	**11.75**	**27.08**	66.75	55.25	53.13

Note Inflated FPR (≥7.5%) are underlined and bolded

Table 2 Averaged FPR (%) and TPR (%) under the conditions with sample sizes of the reference and the focal groups: 800/200

			FPR			TPR		
Impact	DIF type	DIF (%)	OR with purification	OR without purification	CI	OR with purification	OR without purification	CI
Impact = 0		0	4.35	4.05	4.95	–	–	–
	Balanced DIF	10	4.22	4.00	5.11	67.50	68.50	76.00
		20	4.25	4.00	5.06	68.75	69.75	78.25
		30	4.50	4.21	5.43	67.67	68.83	79.50
		40	4.00	3.67	5.33	68.00	68.88	80.00
	Unbalanced DIF	10	4.50	4.33	5.56	65.50	64.50	70.50
		20	4.25	4.56	**8.38**	60.25	55.75	59.75
		30	6.36	6.29	**12.93**	53.50	47.00	49.00
		40	**10.67**	**11.00**	**19.08**	44.88	36.38	38.25
Impact=1		0	3.55	3.45	5.35	–	–	–
	Balanced DIF	10	3.61	3.38	5.11	60.00	60.00	76.00
		20	4.19	3.50	5.06	63.75	63.75	76.25
		30	4.07	3.50	5.36	63.83	64.50	74.67
		40	5.33	3.50	5.67	65.88	67.13	72.38
	Unbalanced DIF	10	4.00	3.78	6.11	53.00	51.50	66.50
		20	5.13	4.75	**7.75**	54.50	50.50	54.50
		30	6.79	6.50	**11.93**	46.83	40.83	44.50
		40	**14.42**	**12.00**	**16.50**	36.75	29.75	32.88

Note Inflated FPR (≥7.5%) are underlined and bolded

Table 3 The intervals of the OR and CI methods for the real data

Item no.	OR	CI
1	(0.195, 0.7438)	(−0.247, 0.372)
2	(0.391, 1.102)	(−0.650, 0.121)
3	**(−0.396, 0.399)**	**(0.196, 1.034)**
4	(0.233, 1.012)	(−0.481, 0.357)
5	**(−0.203, 0.426)**	**(0.138, 0.846)**
6	(0.567, 1.112)	**(−0.738, −0.111)**
7	(0.377, 0.993)	(−0.564, 0.122)
8	**(−0.168, 0.454)**	**(0.111, 0.812)**
9	(0.312, 0.860)	(−0.404, 0.214)
10	(0.484, 1.210)	*(−0.783, −0.001)*
11	(0.296, 0.850)	(−0.396, 0.232)
12	(0.403, 0.993)	(−0.570, 0.100)
13	(0.219, 0.910)	(−0.431, 0.335)
14	(0.374, 0.925)	(−0.494, 0.127)
15	(0.135, 0.736)	(−0.243, 0.426)
16	(0.168, 0.717)	(−0.228, 0.394)
17	(0.459, 1.246)	(−0.798, 0.044)
18	(0.523, 1.235)	*(−0.829, −0.027)*
19	**(−0.261, 0.300)**	**(0.305, 0.942)**
20	(0.125, 0.677)	(−0.180, 0.447)
21	(0.345, 0.888)	(−0.445, 0.166)
22	(0.193, 0.858)	(−0.365, 0.362)
23	**(−0.164, 0.421)**	**(0.158, 0.804)**
24	(−0.339, 1.256)	(−0.770, 0.855)
25	(0.529, 2.395)	**(−1.915, −0.052)**
26	(−0.149, 1.210)	(−0.700, 0.734)

CI method identified not only Items 3, 5, 8, 19 and 23 as DIF items, but also Items 6, 10, 18 and 25. Based on the result from the OR method, the real data could be contaminated with unbalanced DIF items because the intervals of the identified DIF items all fell on the same side of the median. According to the simulation results in Tables 1 and 2, the CI method yielded inflated FPR when test had 20% or more unbalanced DIF items.

To reduce the inflated FPR of the CI method, we proposed a two-stage CI method to detect DIF items, which was implemented as follows. At the first stage, we detected the DIF items by using the CI method. Suppose $\{i_1, i_2, \ldots, i_k\}$ were the collection of the DIF items identified by the CI method. At the second stage, we check, for $j = 1, \ldots, k$, if item i_k is a real DIF item by deleting the other DIF items, and use only item i_k and the other non-DIF items to fit the Rasch model and then to detect

if item i_k is a DIF item based on the CI method again. Based on the two-stage CI method, the identified DIF items were Items 3, 5, 6, 8, 19, 23 and 25, the credible intervals of these items were underlined and bolded in Table 3. Items 10 and 18 were identified as DIF items at the first stage, but were not identified as DIF items at the second stage, and the credible intervals of these two items were marked in *italic* and underlined in Table 3.

5 Concluding Remarks

In this article, we compared the finite sample performance of the CI and OR methods for detecting the need of an additional difficulty parameter for the responses of the focal group when the data follow the Rasch model. Simulation studies showed that the CI method worked better than the OR method under the balanced DIF conditions. However, the CI method yielded inflated FPR under the unbalanced DIF condition. The two methods were then applied to an empirical example. Comparisons of these two methods to other IRT models will be an interesting future line of research.

Acknowledgements The research was supported by Academia Sinica and the Ministry of Science and Technology of the Republic of China under grant number MOST 106-2118-M-001-003-MY2. The authors would like to thank Ms. Yi-Jhen Wu for her helpful comments and suggestions.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: Wiley.

Chang, J., Tsai, H., Su, Y.-H., & Lin, E. M. H. (2016). A three-parameter speeded item response model: estimation and application. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (Vol. 167, pp. 27–38). Switzerland: Springer. https://doi.org/10.1007/978-3-319-38759-8_3.

Chang, Y.-W., Tsai, R.-C., & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika, 79,* 255–274. https://doi.org/10.1007/s11336-013-9336-2.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6,* 269–279. https://doi.org/10.1207/s15324818ame0604_2.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15–26. https://doi.org/10.1177/014662169602000102.

Frederickx, S., Tuerlinckx, F., de Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47,* 432–457. https://doi.org/10.1111/j.1745-3984.2010.00122.x.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67,* 373–393. https://doi.org/10.1177/0013164406294781.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* 647–667.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Jin, K. -Y., Chen, H. -F., & Wang, W. -C. (2018). Using odds ratios to detect differential item functioning. *Applied Psychological Measurement, 42,* 613–629. https://doi.org/10.1177/0146621618762738.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement, 75,* 22–56. https://doi.org/10.1177/0013164414529792.

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

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

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20,* 257–274. https://doi.org/10.1177/014662169602000306.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502. https://doi.org/10.1007/bf02294403.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116. https://doi.org/10.1177/014662169301700201.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194. https://doi.org/10.1007/bf02294572.

Su, Y.-H., Chang, J., & Tsai, H. (2018). Using credible intervals to detect differential item functioning in IRT Models. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology research* (Vol. 233, pp. 297–304). Switzerland: Springer. https://doi.org/10.1007/978-3-319-77249-3_25.