Using Credible Intervals to Detect Differential Item Functioning in IRT Models

Ya-Hui Su, Joyce Chang and Henghsiu Tsai

Abstract Differential item functioning (DIF) occurs when individuals from different groups with the same level of ability have different probabilities of answering an item correctly. In this paper, we develop a Bayesian approach to detect DIF based on the credible intervals within the framework of item response theory models. Our method performed well for both uniform and non-uniform DIF conditions in the two-parameter logistic model. The efficacy of the proposed approach is demonstrated through simulation studies and a real data application.

Keywords Credible interval · DIF · Item response model · Markov chain Monte Carlo

1 Introduction

The unidimensional item response theory (IRT) models are statistical models that describe the relationship among a latent trait (intelligence, ability, attitude, etc.), the properties of items, and how respondents answer individual items. Like other statistical models, checking the validity of these models is necessary for the applicability and the success of interpretation. Differential item functioning (DIF) refers to a strong violation of the assumptions in IRT models. More specifically, DIF occurs when individuals from different groups with the same level of ability have different

Y.-H. Su
Department of Psychology, National Chung Cheng University, 168 University Road,
Section 1, Min-Hsiung, Chia-Yi 62102, Taiwan
e-mail: psyyhs@ccu.edu.tw

J. Chang
Department of Economics, University of Texas at Austin, 2225 Speedway,
Austin, TX 78712, USA
e-mail: joyce.chang@utexas.edu

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

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_25

probabilities of answering an item correctly. Studies of DIF deal with the question of how item scores are affected by external variables that do not belong to the construct to be measured (Glas 1998). Therefore it is important to know which items in a test are subject to DIF.

Many DIF detection methods have been proposed in the literature, including techniques based on the Mantel-Haenszel statistic (Holland and Thayer 1988; Camilli and Penfield 1997; Li 2015), the log-linear models (Kok et al. 1985; Dancer et al. 1994), the IRT models (Hambleton and Rogers 1989; Wang and Woods 2017), and the log-linear IRT models (Kelderman 1989). See Glas (1998) for further discussions. Glas (1998) used the Lagrange multiplier test to evaluate DIF within the framework of several IRT models, including the Rasch model, the one-parameter logistic (1PL), and the two-parameter logistic (2PL) models.

In terms of statistical inference, there are two major approaches: frequentist inference and Bayesian inference. Using the approach of frequentist inference, hypothesis testing and confidence intervals play important roles, and conclusions are drawn based on the frequency or proportion of the observed data. A confidence interval (CI) is a type of interval estimate (of a population parameter) that is computed from the observed data. Confidence intervals (CIs) can be used as a significance test. The simple rule is that if the 95% CI does not include the null value, the null hypothesis is rejected at 0.05 level (e.g., Dahiru 2008, p. 25).

Using the approach of Bayesian inference, a credible interval is an interval in the domain of a posterior probability distribution or a predictive distribution, and is used for interval estimation. See Sect. 7.3 of Garthwaite et al. (2002) for further discussion. So, similar to the frequentist approach, if one uses a Bayesian approach, the null hypothesis is rejected at 0.05 level if the 95% credible interval does not include the null value. Riley and Carle (2012) used 95% credible intervals to assess differences in how respondents answer items administered by computerized adaptive testing versus paper-and-pencil. Nevertheless, their study only focused on uniform DIF without considering non-uniform DIF, and was limited to a small number of replications per experimental condition.

Our goal of this study is to adopt a Bayesian approach to evaluate DIF within the framework of IRT models by using credible intervals. In this paper, we obtained 95% credible intervals to analyze both uniform and non-uniform DIF in the context of 2PL models. The rest of the article is organized as follows. Section 2 introduces our method to detect DIF within the framework of 2PL models. Section 3 describes simulations to investigate the performance of the proposed method in finite samples. Section 4 applies the proposed analysis to the data of the physics examination of the 2010 Department Required Test in Taiwan, and Sect. 5 provides some concluding remarks.

2 Detecting Differential Item Functioning in Two-Parameter Logistic Item Response Model

Let Y_{pj} be the dichotomous response of examinee p on item j, where $p = 1, 2, ..., P$, and $J = 1, 2, ..., J$. Denote b_j and a_j as the location and scale parameters respectively, for item j, and θ_p as the ability parameter for examinee p. In the 2PL model (Birnbaum 1968), the probability of examinee p getting a correct response on item j is given by

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

The parameter a_j is also known as the discrimination parameter (de Ayala 2009), or the slope parameter (Wang 2004), and the parameter b_j is called the difficulty parameter in Embretson and Reise (2000) and Wang and Xu (2015). For more descriptions and discussions of the 2PL model, see Embretson and Reise (2000), Wang (2004), and de Ayala (2009).

An item is said to exhibit DIF if the probability of correctly answering the item differs across separate subgroups after controlling for the underlying ability. Specifically, consider the simplest case of two groups, namely the reference and focal group, and use $g_p = 0$ and $g_p = 1$ to indicate whether the examinee p belongs to the reference group or the focal group. Furthermore, each group has its own difficulty and discrimination parameters. Then, Eq. (1) becomes

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

where a_j and c_j are the discrimination parameters and b_j and d_j are the difficulty parameters for the reference and the focal group, respectively. Alternatively, we can adopt the notations of Glas (1998) to write Eq. (2) as

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

Equation (3) implies that the responses of the reference group are properly described by (1), but that the responses of the focal group need additional difficulty parameters δ_j, additional discrimination parameters γ_j, or both. Therefore, we consider the following two hypotheses:

$$H_{\gamma_j,0} : \gamma_j = 0 \text{ versus } H_{\gamma_j,1} : \gamma_j \neq 0,$$
$$H_{\delta_j,0} : \delta_j = 0 \text{ versus } H_{\delta_j,1} : \delta_j \neq 0.$$

Due to the complexity of the likelihood function, a Bayesian estimation method is often used. Specifically, we follow closely the Bayesian approaches of Chang et al. (2014, 2016). For model identification purpose, the marginal distribution of θ_p is set to be the standard normal.

The procedure for testing the hypotheses runs as follows. Suppose there are J items in the test. For each item, we test $\gamma_j = 0$ and $\delta_j = 0$ separately, and only focus on one item at a time. Let η_j be either γ_j or δ_j. If $\eta_j = \gamma_j$, then $\tilde{\eta}_j = \delta_j$, and vice versa (if $\eta_j = \delta_j$, then $\tilde{\eta}_j = \gamma_j$). Then, a size α test of $\eta_j = 0$ is constructed as follows. First, let item j follow Eq. (3) and set $\tilde{\eta}_j = 0$, whereas the other items follow Eq. (1). In other words, we only focus on testing, if for item j, the responses of the focus group need an additional parameter η_j. Then, we implement the Bayesian analysis via the Markov chain Monte Carlo (MCMC) scheme to construct the equal-tailed $1 - \alpha$ credible interval for the parameter η_j. If the interval includes 0, then we do not reject $\eta_j = 0$. Otherwise, $\eta_j = 0$ is rejected.

3 Simulation Study

In this section, we describe the simulation studies to evaluate the performance of our tests. We fixed the Type-I error of each test (α) to 0.05. All computations were performed using Fortran code with IMSL subroutines. For each p, g_p is randomly assigned to be 0 or 1 with a probability of .50. In each experiment, we simulate a test consisting of 10 items, i.e., $J = 10$. The number of examinees (P) are 200 and 400 students. For the true values of a_j and b_j, for $j = 1, ..., J$, we fit the data of the 26 items of the physics examination (see Sect. 4) to the 2PL model defined in Eq. (1), and use the fitted values of the a_j and the b_j of the first 10 multiple-choice items to be the true values. Regarding the values of γ_j and δ_j, we consider two cases (see Table 1). The first case is that there is only one item with $\gamma_j \neq 0$ or $\delta_j \neq 0$, but not both. The second case is that there are three items of $\gamma_j = 1$ or $\delta_j = 1$, or both. The results are summarized in Table 2.

To construct the credible intervals, we produce 11,000 MCMC draws with the first 1,000 draws as burn-in. For each experiment and each item, we repeat the exercise 1,000 times to create 1,000 credible intervals to get the empirical probability of detecting the DIF. In Table 2, p_η^P is used to denote the probability of rejecting the hypothesis $\eta_j = 0$ for the value of P. Again, η denotes either γ or δ. When a test is used to test $\eta_j = 0$, the probabilities of rejecting the hypothesis $\eta_j = 0$ when it is true and when it is not true are the so-called Type-I error and the power of the test, respectively. In Table 2, the numbers with and without parentheses correspond to power and type-I error, respectively.

As shown in Table 2, it is clear that for DIF items the power increases with the value of P. For non-DIF items the Type-I errors are on average close to the nominal size, although some of them are as large as 0.131 (p_δ^{400} of item 9 for the case of one DIF item) and as small as 0.009 (p_γ^{200} of item 8 for the case of 3 DIF items).

Table 1 Overview of the experiments

Nr. of DIF items	Condition	Test	P
One	1	$\gamma_j = 0$	200, 400
	2	$\delta_j = 0$	200, 400
Three	3	$\gamma_j = 0$; $\delta_j = 0$	200, 400

Table 2 Empirical probabilities of rejecting $\gamma_j = 0$ and those of $\delta_j = 0$

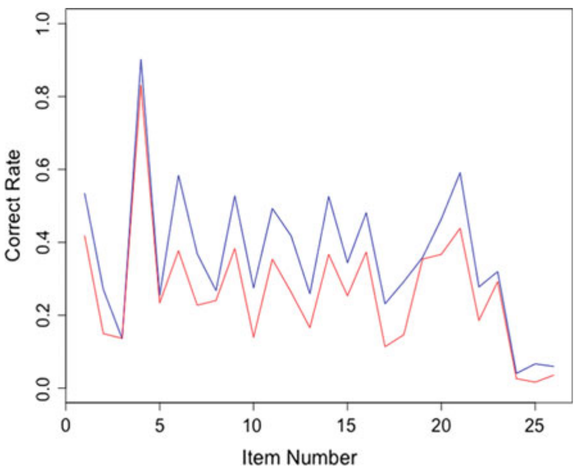
True values			$\gamma_1 = 1$		$\delta_1 = 1$		$\gamma_1 = 1$; $\delta_2 = 1$; $\gamma_3 = 1$ and $\delta_3 = 1$			
Item	a_j	b_j	p_γ^{200}	p_γ^{400}	p_δ^{200}	p_δ^{400}	p_γ^{200}	p_γ^{400}	p_δ^{200}	p_δ^{400}
1	1.195	−0.001	(0.152)	(0.317)	(0.772)	(0.979)	(0.244)	(0.347)	0.117	0.083
2	1.242	1.524	0.042	0.052	0.056	0.058	0.045	0.051	(0.680)	(0.926)
3	0.544	1.955	0.034	0.038	0.058	0.053	(0.121)	(0.222)	(0.149)	(0.249)
4	0.778	−2.195	0.045	0.049	0.103	0.080	0.026	0.048	0.094	0.077
5	0.803	1.254	0.046	0.056	0.039	0.062	0.039	0.051	0.040	0.053
6	0.841	−0.094	0.055	0.053	0.068	0.062	0.053	0.049	0.065	0.067
7	1.011	0.877	0.046	0.056	0.063	0.075	0.053	0.048	0.058	0.068
8	0.082	1.054	0.070	0.012	0.046	0.060	0.009	0.014	0.048	0.061
9	1.444	0.084	0.042	0.052	0.097	0.131	0.060	0.049	0.077	0.105
10	1.934	1.879	0.055	0.080	0.049	0.057	0.023	0.059	0.047	0.043

Moreover, the Type-I error and the power of the test of $\gamma_j = 0$ do not differ much for one or three DIF items. For the test of $\delta_j = 0$, the Type-I error does not change much for one or three DIF items, whereas the power deteriorates from one to three DIF items. It is also interesting to note that the power of detecting DIF on the difficulty parameter is much larger than that on the discrimination parameter.

4 Application

In this section, the proposed procedure described in the previous sections are 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). Examinees have to answer 26 questions in 80 min. The 26 questions are further divided into three parts. The totel score is 100, and the test is administered under formula-scoring directions. For the first part, there are 20 multiple-choice questions, and the examinees have to choose one correct answer out of 5 possible choices. For each correct answer, 3 points are granted, and 3/4 point is deducted from the raw score for each incorrect answer. The second part consists of 4 multiple-response questions, and each question consists of 5 choices, examinees need to select all the answer choices that apply. The choices in each item are knowledge-related, but are

Fig. 1 Plots of the correct rates for male (blue line) and female (red line) for all items in the test



answered and graded separately. For each correct choice, 1 point is earned, and for each incorrect choice 1 point is deducted from the raw score. The adjusted score would only be 0 or above for each of these two parts. The last part consists of 2 calculation problems, and deserves 20 points in total.

The data from 1,000 randomly sampled examinees contains the original responses and nonresponses information, but we treat both nonresponses and incorrect answers the same way and code them as $Y_{pj} = 0$ as suggested by Chang et al. (2014). As for the calculation part, the response Y_{pj} is coded as 1 whenever the original score is more than 7.5 out of 10 points, and zero otherwise (see also Chang et al. 2014). Chang et al. (2016) showed that the 2PL model fits the data well. Here, we consider male as the reference group, and female as the focal group and among the 1,000 examinees, 692 of them are male and 308 are female.

We make more MCMC draws than in Sect. 3. Specifically, we produce 40,000 MCMC draws with the first 10,000 draws as burn-in. Then we test $\gamma_j = 0$ and $\delta_j = 0$, for $j = 1, ..., 26$. Again, we consider $\alpha = 0.05$. The results show that for Item 6, the discrimination and the difficulty parameters are both subject to DIF, whereas for Item 24, only the discrimination parameter is subject to DIF, and for items 7, 17, 18, and 21, only the difficulty parameter is subject to DIF. To further study the testing results, we first note that for each item, and for each examinee, the score can either be 0 or 1. Therefore, for each item, we define the percent of correct rate of each gender to be the percent of scoring 1. The results are summarized in Fig. 1. It is interesting to note that the correct rates for the male are all higher than those for the female, except for items 3 and 19. For these two items, they are almost identical.

Then, we plot the credible intervals for the γ and the δ parameters in Fig. 2. In this figure, the dot in the middle of each interval represents the median of the posterior distribution based on the MCMC draws after burn-in. For the two items the discrimination parameter is subject to DIF: for Item 6, the discrimination parameter is higher for females than for males; for Item 24, the opposite holds.

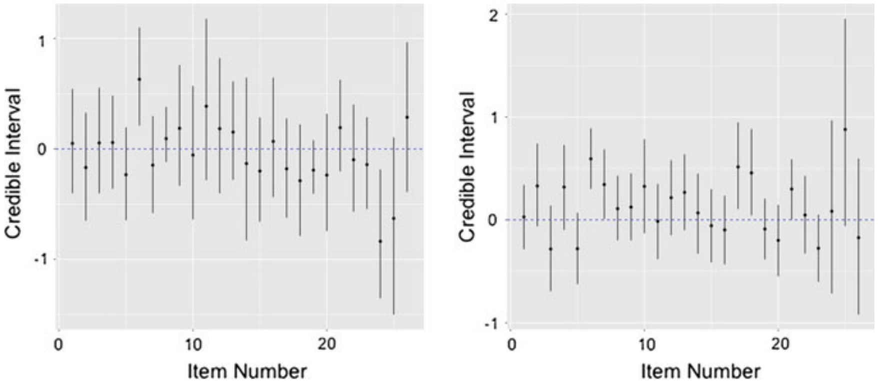


Fig. 2 Plots of the credible intervals for all items for γ_j (left figure) and δ_j (right figure)

From Fig. 1, we know that Item 6 is a relatively easy item and Item 24 is a relatively difficult item. For the 5 items that the difficulty parameter are subject to DIF, it is always that the parameter for the female is higher than that for the male. The results are consistent with Fig. 1.

5 Concluding Remarks

In this article, we propose to use credible intervals to detect DIF in 2PL models. Simulation studies show that the proposed method works reasonably well for detecting the need of an additional difficulty parameter or an discrimination parameter for the responses of the focus group. Applications of the proposed method to other IRT models will be an interesting future line of research. It will also be worthwhile to compare the power of our test with others in the future.

Acknowledgements The research was supported by Academia Sinica and the Ministry of Science and Technology of the Republic of China under grant number MOST 104-2118-M-001-008-MY2. The authors would like to thank Rianne Janssen, the Co-Editor, Dr. Yu-Wei Chang and Ms. Yi-Jhen Wu for their helpful comments and suggestions.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based ore Mantel-Haenszel statistics. *Journal of Educational Measurement*, *34*, 123–139.

Chang, Y.-W., Tsai, R.-C., & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika*, *79*, 255–274.

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.

Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan Postgraduate Medicine*, *6*, 21–26.

Dancer, L. S., Anderson, A. J., & Derlin, R. L. (1994). Use of log-linear models for assessing differential item functioning in a measure of psychological functioning. *Journal of Consulting and Clinical Psychology*, *62*, 710–717.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: L. Erlbaum Associates.

Garthwaite, P., Jolliffe, I., & Jones, B. (2002). *Statistical inference*. Oxford: Oxford University Press.

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

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*, 313–334.

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

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika*, *54*, 681–697.

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement*, *22*, 295–303.

Li, Z. (2015). A power formula for the Mantel-Haenszel test for differential item functioning. *Applied Psychological Measurement*, *39*, 373–388.

Riley, B. B., & Carle, A. C. (2012). Comparison of two Bayesian methods to detect mode effects between paper-based and computerized adaptive assessments: A preliminary monte carlo study. *BMC Medical Research Methodology*, *12*, 124.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*, 456–477.

Wang, M., & Woods, C. M. (2017). Anchor selection using the Wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, *41*, 17–29.

Wang, W.-C. (2004). Rasch measurement theory and application in education and psychology. *Journal of Education and Psychology*, *27*, 637–694. (in Chinese).