Use of Information Criteria in the Study of Group Differences in Trace Lines

Applied Psychological Measurement
2019, Vol. 43(2) 95–112
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618772292
journals.sagepub.com/home/apm

Seock-Ho Kim¹, Allan S. Cohen¹, Sun-Joo Cho² and Hyo Jin Eom³

Abstract

A brief review of various information criteria is presented for the detection of differential item functioning (DIF) under item response theory (IRT). An illustration of using information criteria for model selection as well as results with simulated data are presented and contrasted with the IRT likelihood ratio (LR) DIF detection method. Use of information criteria for general IRT model selection is discussed.

Keywords

Akaike information criterion, Bayesian information criterion, differential item functioning, information criteria, item response theory, likelihood ratio test, marginal maximum likelihood estimation

Introduction

According to Rao and Wu (2001), almost all statistical problems can be considered as model selection problems. Detection of differential item functioning (DIF), in fact, can be viewed as a model selection problem where the augmented model that allows for groups that differ in their answer patterns to an item is compared with the compact model that allows for groups that exhibit no differences in their answer patterns. Although the likelihood ratio (LR) test (Rao, 1973) may provide an ideal means to compare model parameters, it may not be an optimum, when used with latent variable modeling including item response theory (IRT). This is partly due to sparseness of data, having too many possible item response patterns to observe relative to sample size (Bock & Moustaki, 2007). The authors applied information criteria for model selection to the detection of DIF under IRT in this study because model selection based on information theory approaches may provide a new paradigm that is different from the usual hypothesis testing methods (Burnham & Anderson, 1998).

There are other related studies that used some information criteria for model comparison including the detection of DIF. For example, Klein Entink, Fox, and van der Linden (2009) used one of the information criteria called the deviance information criterion (DIC) by Spiegelhalter, Best, Carlin, and van der Linde (2002) in the model selection for the test speededness model

¹University of Georgia, Athens, USA

²Vanderbilt University, Nashville, TN, USA

³Korea University, Seoul, Korea

Corresponding Author:

Seock-Ho Kim, Department of Educational Psychology, University of Georgia, 325 Aderhold Hall, Athens, GA 30602-7143, USA.

Email: shkim@uga.edu

with a multivariate multilevel structure via Markov chain Monte Carlo estimation. Magis, Tuerlinckx, and De Boeck (2015) proposed the logistic regression lasso approach for the detection of DIF and contrasted that with several information criteria (e.g., the weighted information criterion). May (2006) used both the LR chi-square test based on the deviance statistic and the Bayesian information criterion (BIC) for model fit comparison in conjunction with a multilevel Bayesian IRT method. Tutz and Schauberer (2015) proposed the use of the lasso penalty for the detection of DIF and considered cross-validation and information criteria under the Rasch models. None of these papers, however, employed the same method and the framework presented in the current study.

As a didactic article, a brief review of various information criteria is presented below. Full descriptions of the information criteria with detailed theoretical explications can be found in the respective references cited. At the outset, it should be mentioned that the information criteria and the detection of DIF are presented in the context of marginal maximum likelihood estimation of item parameters (Bock & Aitkin, 1981; Bock & Lieberman, 1970). Various information criteria have been suggested and employed recently in conjunction with IRT model selection (e.g., de Ayala, 2009; Kang & Cohen, 2007; Kang, Cohen, & Sung, 2009; Li, Cohen, Kim, & Cho, 2009). Cohen and Cho (2016) presented a detailed review of information criteria and a thorough summary of studies that used information criteria for model selection in IRT (see Cohen & Cho, 2016, Table 18.1).

Note that DIF and its detection methods are well understood in general so will not be presented in detail. Holland and Wainer (1993), for example, discussed a number of issues and various methods for the detection of DIF. Lord (1980) and Thissen, Steinberg, and Gerrard (1986), respectively, described the IRT-Wald test and the IRT likelihood ratio (IRT-LR) test for detection of DIF as used in this study. The trace line or item characteristic curve under the three-parameter logistic (3PL) model (Lord, 1980) is defined as

$$T = c + \frac{1 - c}{1 + \exp[-Da(\theta - b)]}, \quad (1)$$

where θ is ability, $\xi = \{a, b, c\}$ is the set of item parameters, a (slope/discrimination), b (location/difficulty), and c (lower asymptote/guessing), and D is so-called the scaling constant that is 1.7 in this study. The trace lines from the focal and reference groups are compared in a DIF study. For the IRT-Wald test, for example, the two trace lines can be compared for only location parameters with constraints of $a_F = a_R$ and $c_F = c_R$ using a test statistic that is distributed as the one degree of freedom chi-square random variable,

$$X^2(1) = \frac{(b_R - b_F)^2}{SE^2(b_R) + SE^2(b_F)}, \quad (2)$$

where the subscripts R and F designate the reference group and the focal group, respectively, and SE s are the standard errors. Note that a short description of the IRT-LR test is presented in later sections.

Brief Review of Information Criteria

Akaike (1973) presented an information criterion (Akaike information criterion [AIC]) to assess the goodness of fit of a model to data (see deLeeuw, 1992; Parzen, Tanabe, & Kitagawa, 1998) that is based on the expected Kullback–Leibler distance (Burnham & Anderson, 1998; Kullback, 1959; Judge, Griffiths, Hill, Lütkepohl, & Lee, 1985). AIC is generally defined as

$$\text{AIC}(\hat{\theta}) = -2 \log L(\hat{\theta}) + 2k, \quad (3)$$

where $L(\hat{\theta}) = p(y|\hat{\theta})$ is the likelihood function for which θ designates the set of parameters, its hat indicates the maximum likelihood estimates, y designates data, and k is the number of parameters (Akaike, 1973; Sakamoto, Ishiguro, & Kitagawa, 1986). In model selection or comparison, a model with a smaller value of AIC is preferred.

Schwartz (1978) suggested selection of a model for which the term, $\log L(\hat{\theta}) - 1/2 \log N$, is largest, where N is the sample size. Schwarz's term is now modified as the BIC:

$$\text{BIC}(\hat{\theta}) = -2 \log L(\hat{\theta}) + k \log N. \quad (4)$$

Akaike (1976) also presented a different but related BIC. As was the case for AIC, a model with a smaller value of BIC is preferred. AIC and BIC are the most frequently used information criteria in IRT.

Other information criteria sometimes reported in IRT-related computer programs include CAIC (consistent AIC) and sample-size-adjusted BIC (SABIC, a.k.a. Rissanen information criterion [RIC]). CAIC (Bozdogan, 1987) is defined as

$$\text{CAIC}(\hat{\theta}) = -2 \log L(\hat{\theta}) + k \log N + k. \quad (5)$$

Fujikoshi and Satoh (1997) also used the same name of CAIC to designate the corrected AIC following Sugiura (1978) and Hurvich and Tsai (1989). RIC by Rissanen (1978; Sclove, 1987) is defined as

$$\text{RIC}(\hat{\theta}) = -2 \log L(\hat{\theta}) + k \log \frac{N+2}{24}. \quad (6)$$

Hurvich and Tsai (1989) suggested AIC_c for small-sample, second-order, bias adjustment (cf. c-AIC, corrected-AIC, by Sugiura, 1978),

$$\text{AIC}_c(\hat{\theta}) = -2 \log L(\hat{\theta}) + k \left(\frac{2N}{N-k-1} \right), \quad (7)$$

which is not the equation presented in their paper but modified to the current context (see Burnham & Anderson, 1998). Burnham and Anderson (1998) indicated that AIC_c can be useful for, say, $N/k < 40$.

According to Burnham and Anderson (1998), Takeuchi information criterion (TIC) is defined as

$$\text{TIC}(\hat{\theta}) = -2 \log L(\hat{\theta}) + 2 \text{tr} \left(\hat{J}(\theta_o) [\hat{I}(\theta_o)]^{-1} \right) \quad (8)$$

(see Burnham & Anderson, 1998, for the exact definitions of the terms in the above equation). Shibata (1989) recommended an estimator for the trace term that is k (i.e., AIC). In addition, in conjunction with Markov chain Monte Carlo estimation, DIC was developed by Spiegelhalter et al. (2002) but is not employed in the current study (see, for example, Klein Entink et al., 2009). In a sense, all information criteria are measures of badness of fit for which a model with a smaller value of respective information criteria is preferred.

Method

For illustrative purposes, computations of the log likelihood functions and a set of information criteria for the detection of DIF were obtained using one anchor item and one candidate item from a data set from Thissen, Steinberg, and Wainer (1988). The computer program MULTILOG (Thissen, Chen, & Bock, 2002) that implemented marginal maximum likelihood estimation of item parameters (Bock & Aitkin, 1981; Bock & Lieberman, 1970) was used. Note that IRTPRO (Cai, Thissen, & du Toit, 2010) and flexMIRT (Cai, 2012) can also be used to obtain similar parameter estimates.

Because the likelihood function plays an important role in both parameter estimation and computation of all information criteria, the definition of the likelihood function and other relevant equations under the one-parameter logistic (1PL) model (i.e., Thissen, 1982) with data consist of response patterns in conjunction with marginal maximum likelihood estimation for multiple groups may be helpful. These are presented in detail in, for example, Baker and Kim (2004). For the patterned data from multiple groups, it is also possible to obtain the deviance G^2 from the logarithm of a ratio of likelihoods (i.e., the LR statistic; Bishop, Fienberg, & Holland, 1975; Rao, 1973), one from the current model and the other from the saturated model. For example,

$$G^2 = \sum_{g=1}^G \left[-2 \sum_{l=1}^L r_{gl} \log \hat{P}_{gl} - \left(-2 \sum_{l=1}^L r_{gl} \log \frac{r_{gl}}{N_g} \right) \right] = 2 \sum_{g=1}^G \sum_{l=1}^L r_{gl} \log \frac{r_{gl}}{N_g \hat{P}_{gl}}, \quad (9)$$

where $\hat{P}_{gl}$ is the unconditional probability of the response pattern x_{gl} from which the ability parameter from a randomly sampled person θ with the ability distribution conditioned upon the hyperparameters of the group has been integrated out and $p_{gl} = r_{gl}/N_g$ is the observed proportion with r_{gl} is the number of occurrences of pattern l in group g and N_g is the sample size of group g (see Baker & Kim, 2004).

For the IRT-LR test, the negative two log likelihood of item parameters for which the ability has been integrated out, $-2 \log L(\hat{\xi})$, is obtained for the compact model and for the augmented model separately. The augmented model contains more parameters to estimate than the compact model; hence, $-2 \log L(\text{Compact}) > -2 \log L(\text{Augmented})$. The difference between the two models is the IRT-LR test statistic $G^2(df)$ with the degree(s) of freedom df that reflects the difference between the numbers of parameters of the two models.

To illustrate further, results from the simulated data of Thissen et al. (1988) for four different data sets for the detection of DIF were compared with the results from information criteria: (a) one-item anchor under the 1PL model, (b) 50-item anchor under the 1PL model, (c) 50-item anchor under the 3PL model, and (d) four-item anchor under the 3PL model.

In addition, to show a unified approach to the IRT model selection and the detection of DIF using information criteria, an example data set, G341-19.DAT, from Thissen (2001) was calibrated with 1PL, two-parameter logistic (2PL), 3PL, and 3PL-C (i.e., 3PL with a common lower asymptote c) models for the model selection and employed under the 3PL and 3PL-C models for the DIF detection. MULTILOG (Thissen et al., 2002) was used in the calibration runs and the detection of DIF with the IRT-LR test. G341-19.DAT contains item responses of 2,844 examinees from two groups to the dichotomously scored 12 items. Full descriptions of the data can be found in Thissen (2001). Note that trivially different DIF detection results were obtained and presented in Thissen (2001) under the 3PL model with the computer program IRTLRDIF.

Finally, portions of simulation results from Cohen, Kim, and Wollack (1996) were used to obtain empirical error rates of information criteria. In that study, the null data without any DIF were generated for two sample sizes of 500 (i.e., 250 focal and 250 reference groups) and 2,000

(i.e., 1,000 focal and 1,000 reference groups) under the 3PL model with the scaling constant $D=1.7$. The type I error rate of the IRT-LR test was investigated with the null situation. Their study contained type I error results for 2PL, 3PL-C, and 3PL. Only the 3PL portions were used in the current study. The simulation results were presented with 50 replications of the DIF detection cases for the two sample sizes.

There were 50 items on the test, and the parameters used to generate data were originally reported by Lord (1968). The computer program MULTILOG was used to obtain item parameter estimates, the latent group parameter estimates, and the log likelihood values. The IRT-LR test was performed, and the calculations of AIC and BIC as well as other information criteria results were based on the G^2 values from the Cohen et al.'s study. The degree of freedom for the G^2 was three, and the difference in the parameters between the compact model and the augmented model for AIC and BIC was also three.

Results

An Illustration

Table 1 presents the data and calibration results for the one anchor item and the one candidate item from Thissen et al. (1988). There are two items for the two groups, hence, there are four response patterns for the reference group and four response patterns for the focal groups. Table 1 shows two 2×2 tables, one for each of the two groups in which each cell contains a combination of the correct response 1 and the incorrect response 0 to the anchor and candidate items. There are 2,000 examinees in the reference group and 500 examinees in the focal group.

According to the simulation condition, the underlying ability distribution for the reference group was $N(0, 1)$ and the focal group was $N(-1, 1)$. The item parameters were expressed on the so-called normal ogive metric (i.e., $D=1.7$); the item difficulty parameter for the anchor item was $b=0.04$, and the item difficulty parameters for the candidate item were $b_R = -1$ and $b_F = -0.7$ for the reference group and the focal group, respectively.

The rest of Table 1 shows the calibration results under the unconstrained model and the constrained model. The unconstrained model is the augmented model where $b_R \neq b_F$, and the constrained model is the compact model where $b_R = b_F$ in conjunction with the IRT-LR test (Thissen et al., 1986). The IRT-Wald test and the IRT-LR test also were performed and the results included in Table 1. Both the IRT-Wald test and the IRT-LR test were statistically significant at the nominal $\alpha = .05$ level indicating the candidate item shows DIF. Note that the calibration results in Table 1 were obtained from the computer program MULTILOG (Thissen et al., 2002) and nearly the same as those presented in Table 10.1 of Thissen et al. (1988).

Information Criteria

Table 2 contains the values of $-2 \log$ likelihood, the number of parameters estimated, and the values of six information criteria for the augmented model and for the compact model, respectively. Note that θ used in the earlier equations of information criteria has been replaced with ξ because the likelihood is marginalized with regard to θ under the marginal maximum likelihood estimation. The model with the smaller value of the information criterion is preferred. The values of $-2 \log$ likelihood for the augmented and compact models were calculated with short code in R (Venables, Smith, & The Development Core Team, 2009) from the parameter estimates from the MULTILOG calibration runs and the original data. AIC, RIC, AIC_c, and TIC supported the augmented model, and BIC and CAIC supported the compact model. Hence the

Table 1. Data and Calibration Results for One Anchor Item and One Candidate Item.

Data							
Reference group				Focal group			
		Candidate Item				Candidate Item	
		0	1			0	1
Anchor Item	0	368	661	Anchor Item	0	278	125
	1	91	880		1	27	70
True parameters:		$\mu_R = 0$				$\mu_F = -1$	
(Candidate item:)		$b_R = -1$				$b_F = -0.7$	
(Anchor item:)					$b = 0.04$		
					$a = 1$		
					$\sigma = 1$		
		Not constrained, $b_R \neq b_F$					
MML estimates:		$\mu_R = 0$ (fixed)				$\mu_F = -1.10$ (0.06)	
(Candidate item:)		$b_R = -1.00$ (0.05)				$b_F = -0.72$ (0.08)	
(Anchor item:)					$b = 0.05$ (0.04)		
					$a = 1.12$ (0.05)		
					$\sigma = 1$ (fixed)		
$G^2(3) = 0.143$							
(Approximate standard errors in parentheses)							
$-2 \log L(\text{Augmented}) = 5,822.830$						$-2 \log L(\text{Saturated}) = 5,822.687$	
IRT-Wald test, $b_R = b_F$: $X^2(1) = (-1.00 - (-0.72))^2 / 0.0089 = 8.346$							
		Constrained, $b_R = b_F$					
MML estimates:		$\mu_R = 0$ (fixed)				$\mu_F = -1.27$ (0.06)	
(Candidate item:)		$b_R = -0.97$ (0.04)				$b_F = -0.97$ (equal)	
(Anchor item:)					$b = 0.03$ (0.04)		
					$a = 1.11$ (0.04)		
					$\sigma = 1$ (fixed)		
$G^2(4) = 5.665$							
(Approximate standard errors in parentheses)							
$-2 \log L(\text{Compact}) = 5,828.352$						$-2 \log L(\text{Saturated}) = 5,822.687$	
IRT-LR test, $b_R = b_F$: $G^2(1) = 5,828.352 - 5,822.830 = 5.522 \equiv G^2(4) - G^2(3)$							

Note. IRT = item response theory; MML = marginal maximum likelihood.

information criteria yielded inconsistent model selection results. Note that AIC tends to support a more complicated model whereas BIC tends to support a simpler model.

Results With Simulated Data

The results from the four simulation data sets are summarized based on the tables from Thissen et al. (1988). It should be noted that data sets were not recalibrated by any computer programs but rather IRT-LR G^2 values from the original tables for the DIF detection were utilized to obtain results from the information criteria for model selection.

One-item anchor IPL data. Table 3 shows the summary of the results of IRT-LR DIF analysis and information criteria with AIC and BIC from the one-item anchor IPL simulated data. The anchor item of the simulated data had $b = 0.04$ as was used in the illustration shown in Table 1. There were nine cases and three levels of b_R (-1 , 0 , and 1). For each level of b_R , there were three values of b_F that differed from b_R by 0.05 , 0.15 , and 0.30 . For all items $a = 1$ and $c = 0$. Table 3 shows that two of the three b -differences of 0.30 were detected at $\alpha = .05$ with the IRT-LR $G^2(1)$. The b -differences of 0.05 and 0.15 were never detected for DIF. Based on the results from Thissen and Wainer (1982, Table 1), the standard error for b with a sample size of 500 for

Table 2. Values of $-2 \log L(\hat{\xi})$, the Number of Parameters, Information Criteria From the Augmented and Compact Models of the One Anchor Item and One Candidate Item Data With $N=2,500$.

Model	$-2 \log L(\hat{\xi})$	Information criteria						k
		AIC	BIC	CAIC	RIC	AIC _c	TIC ^a	
Augmented	5,822.830	5,832.830	5,861.950	5,866.950	5,846.064	5,832.854	5,832.830	5
Compact	5,828.352	5,836.352	5,859.648	5,863.648	5,846.939	5,836.368	5,836.352	4

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = consistent AIC; RIC = Rissanen information criterion; TIC = Takeuchi information criterion.

^aTIC with the trace term k (Shibata, 1989).

Table 3. One-Item Anchor IPL IRT-LR DIF Analysis Results and Information Criteria (AIC, BIC) Results.

Item	Reference group		Focal group		G^2	Information criteria	
	True b_R	Estimated b_R	True b_F	Estimated b_F		AIC	BIC
4	-1.00	-1.04	-0.95	-1.14	1.2	Compact	Compact
5	-1.00	-1.02	-0.85	-1.03	0.0	Compact	Compact
6	-1.00	-0.99	-0.70	-0.72	5.6*	Augmented	Compact
7	0.00	-0.03	0.05	0.00	0.0	Compact	Compact
8	0.00	0.03	0.15	0.02	0.0	Compact	Compact
9	0.00	0.04	0.30	0.11	0.2	Compact	Compact
10	1.00	1.00	1.05	0.78	2.4	Augmented	Compact
11	1.00	1.04	1.15	0.97	0.3	Compact	Compact
12	1.00	0.99	1.30	1.36	5.0*	Augmented	Compact

Note. IPL = one-parameter logistic; IRT = item response theory; IRT-LR = IRT likelihood ratio; DIF = differential item functioning; AIC = Akaike information criterion; BIC = Bayesian information criterion.

which the mean of the ability distribution is matched with b was 0.07 (i.e., 0.065561513). Hence, the values of b -difference, 0.05 and 0.15, might be difficult to detect in general.

The IRT-LR test statistic $G^2(1)$ is reported for each item in Table 3. The value of $G^2(1)$ is based on the difference between $-2 \log$ likelihood from the compact model and $-2 \log$ likelihood from the augmented model. When the compact model has the number of parameters k , the augmented model has the number of parameters $k + 1$. Therefore, if the value of $G^2(1)$ is greater than 2, then the value of AIC for the augmented model is smaller than that for the compact model (i.e., the augmented model is selected when $G^2(1) \geq 2$ by AIC). Using the same logic, if $G^2(1)$ is greater than $\log N$ or $\log 2500 = 7.824$, the value of BIC for the augmented model is smaller than that for the compact model (i.e., the augmented model is selected when $G^2(1) \geq 7.824$ by BIC).

Again, because $G^2(1) \geq 2$ supports the augmented model by AIC, one case for the b -difference of 0.05 and two cases for the b -difference of 0.30 supported the augmented model by AIC. Because $G^2(1) \geq 7.824$ supports the augmented model by BIC, all nine cases supported the compact model by BIC.

50-item anchor IPL data. Table 4 shows the summary of the results of IRT-LR DIF analysis and information criteria with AIC and BIC from the 50-item anchor 1PL simulated data. The

Table 4. 50-Item Anchor IPL IRT-LR DIF Analysis Results and Information Criteria (AIC, BIC) Results.

Item	Reference group		Focal group		G^2	Information criteria	
	True b_R	Estimated b_R	True b_F	Estimated b_F		AIC	BIC
4	-1.00	-1.02	-0.95	-1.01	0.0	Compact	Compact
5	-1.00	-0.98	-0.85	-0.88	1.9	Compact	Compact
6	-1.00	-1.03	-0.70	-0.62	29.8**	Augmented	Augmented
7	0.00	-0.01	0.05	0.13	2.9	Augmented	Compact
8	0.00	0.04	0.15	0.16	1.8	Compact	Compact
9	0.00	0.05	0.30	0.23	4.8*	Augmented	Compact
10	1.00	0.97	1.05	0.89	0.4	Compact	Compact
11	1.00	1.04	1.15	1.12	0.4	Compact	Compact
12	1.00	0.99	1.30	1.52	14.8**	Augmented	Augmented

Note. IPL = one-parameter logistic; IRT = item response theory; IRT-LR = IRT likelihood ratio; DIF = differential item functioning; AIC = Akaike information criterion; BIC = Bayesian information criterion.

* $p < .05$. ** $p < .01$.

anchor items of the simulated data had equally spaced b values from -2 to 2 with $a = 1$ and $c = 0$. Because the structure of the data was the same, there were nine cases for which three levels of b_R (-1 , 0 , and 1) were crossed with the three levels of b -difference of 0.05 , 0.15 , and 0.30 . Again, for all items, $a = 1$ and $c = 0$. Table 4 shows all the three b -difference of 0.30 were detected at $\alpha = .05$ with the IRT-LR $G^2(1)$; two of them were detected at $\alpha = .01$. The b -differences of 0.05 and 0.15 were never detected for DIF.

Because $G^2(1) \geq 2$ supports the augmented model by AIC, one case for the b -difference of 0.05 and all three cases for the b -difference of 0.30 supported the augmented model by AIC. Because $G^2(1) \geq 7.824$ supports the augmented model by BIC, two of the three b -difference of 0.30 supported the augmented model by BIC.

50-item anchor 3PL data. Table 5 shows the results of IRT-LR DIF analysis and information criteria with AIC and BIC from the 50-item anchor 3PL data. The 50 anchor items had sets of the a and b true parameters randomly sampled from a bivariate distribution and the same true c parameter of $.20$. Item parameters were estimated with the method of maximum a posteriori (i.e., Bayes modal estimation) with a normal prior with hyperparameters of mean -1.4 and standard deviation 0.5 , that is $N(-1.4, 0.5)$, imposed on the logit of c (see Thissen, 1991). The population θ distribution of the reference group was $N(0, 1)$ and that of the focal group was $N(-1, 1)$.

The true parameters for Items 4 to 12 were the same as Items 4 to 12 of the Tables 3 and 4. Because of the values of $c = 0$ and the items were estimated with the normal prior of $\text{logit}(c)$, the prior misspecification was in part considered for Items 4 to 12. The top portions of Table 5 show all items of the b -difference of 0.30 were detected at $\alpha = .05$ and two of them at $\alpha = .01$ with the IRT-LR $G^2(1)$.

Because $G^2(1) \geq 2$ supports the augmented model by AIC, two cases for the b -difference of 0.15 and three cases for the b -difference of 0.30 supported the augmented model by AIC for Items 4 to 12. Because $G^2(1) \geq 7.824$ supports the augmented model by BIC, two cases of candidate Items 6 and 12 for the b -difference of 0.30 supported the augmented model by BIC for Items 4 to 12.

The true parameters for Items 13 to 21 were similar to Items 4 to 12 with regard to a and b , but different in terms of c . Because of the true parameter values of $c = .20$ and items were

Table 5. 50-Item Anchor 3PL IRT-LR DIF Analysis Results and Information Criteria (AIC, BIC) Results.

Item	Reference group		Focal group		Estimated b_F	Information criteria	
	True b_R	Estimated b_R	True b_F	G^2		AIC	BIC
4	-1.00	-0.85	-0.95	-0.84	0.0	Compact	Compact
5	-1.00	-0.86	-0.85	-0.72	2.4	Augmented	Compact
6	-1.00	-0.88	-0.70	-0.61	9.8**	Augmented	Augmented
7	0.00	0.11	0.05	0.21	1.7	Compact	Compact
8	0.00	0.06	0.15	0.22	2.9	Augmented	Compact
9	0.00	0.02	0.30	0.23	5.1*	Augmented	Compact
10	1.00	1.00	1.05	0.95	0.3	Compact	Compact
11	1.00	1.14	1.15	1.32	1.5	Compact	Compact
12	1.00	1.10	1.30	1.79	14.1**	Augmented	Augmented
(For Items 4-12, $a = 1$ and $c = 0$ for both groups)							
13	-1.00	-1.18	-0.95	-1.01	2.4	Augmented	Compact
14	-1.00	-0.98	-0.85	-0.77	4.4*	Augmented	Compact
15	-1.00	-1.14	-0.70	-0.68	17.3**	Augmented	Augmented
16	0.00	-0.04	0.05	0.13	1.8	Compact	Compact
17	0.00	-0.06	0.15	0.13	2.0	Augmented	Compact
18	0.00	0.12	0.30	0.32	2.5	Augmented	Compact
19	1.00	1.12	1.05	0.84	2.0	Augmented	Compact
20	1.00	1.05	1.15	1.18	0.3	Compact	Compact
21	1.00	1.06	1.30	1.55	4.3*	Augmented	Compact
(For Items 15-21, $a = 1$ and $c = .2$ for both groups)							
Item	Reference group		Focal group		Estimated c_F	Information criteria	
	True c_R	Estimated c_R	True c_F	AIC		AIC	BIC
22	.20	.16	.15	.19	1.4	Compact	Compact
23	.20	.22	.10	.13	13.6**	Augmented	Augmented
24	.20	.17	.05	.08	12.7**	Augmented	Augmented
(For Items 22-24, $a = 1$ and $b = 0.5$ for both groups)							

Note. 3PL = three-parameter logistic; IRT = item response theory; IRT-LR = IRT likelihood ratio; DIF = differential item functioning; AIC = Akaike information criterion; BIC = Bayesian information criterion.

* $p < .05$. ** $p < .01$.

estimated with the normal prior of $\text{logit}(c)$ of $N(-1.4, 0.5)$, estimation was done with an ideal prior specification for c for Items 13 to 21. The middle portions of Table 5 show one item of the b -difference of 0.15 and two items of the b -difference of 0.30 were detected at $\alpha = .05$, and Item 15 of the b -difference of 0.30 was detected at $\alpha = .01$ with the IRT-LR $G^2(1)$.

Because $G^2(1) \geq 2$ supports the augmented model by AIC, two cases for the b -difference of 0.05, two cases for the b -difference of 0.15, and all three cases for the b -difference of 0.30 supported the augmented model by AIC for Items 13 to 21. Because $G^2(1) \geq 7.824$ supports the augmented model by BIC, just one case for the b -difference of 0.30 supported the augmented model by BIC for Items 13 to 21.

For Items 22 to 24, DIF was tested with the IRT-LR $G^2(1)$ for $c_R = c_F$. The reference and focal groups shared the same $a = 1$ and $b = 0.5$, but different c s. There were three values of c_F lower than c_R by .05, .10, and .15. Two larger sizes of the c -difference were detected at $\alpha = .01$ with the IRT-LR $G^2(1)$. Based on the results from Thissen and Wainer (1982, Table 5), the standard error for c with a sample size of 500 for which the mean of the ability distribution is matched with b is 0.07 (i.e., 0.066053448). Hence, the values of c -difference, .05 and .10, might

Table 6. Four-Item Anchor 3PL IRT LR-DIF Analysis Results and Information Criteria (AIC, BIC) Results.

Item	Reference group		Focal group		G^2	Information criteria	
	True b_R	Estimated b_R	True b_F	Estimated b_F		AIC	BIC
6	-1.00	-0.79	-0.70	-0.44	7.1**	Augmented	Compact
9	0.00	0.10	0.30	0.48	7.8**	Augmented	Compact
12	1.00	0.98	1.30	1.82	11.4**	Augmented	Augmented
(For Items 6-12, $a = 1$ and $c = 0$ for both groups)							
15	-1.00	-1.20	-0.70	-0.69	10.5**	Augmented	Augmented
18	0.00	0.10	0.30	0.49	4.8*	Augmented	Compact
21	1.00	0.99	1.30	1.54	3.9*	Augmented	Compact
(For Items 15-21, $a = 1$ and $c = .2$ for both groups)							
Item	Reference group		Focal group		G^2	Information criteria	
	True c_R	Estimated c_R	True c_F	Estimated c_F		AIC	BIC
24	.20	.16	.05	.09	3.3	Augmented	Compact
(For Item 24, $a = 1$ and $b = 0.5$ for both groups)							

Note. 3PL = three-parameter logistic; IRT = item response theory; IRT-LR = IRT likelihood ratio; DIF = differential item functioning; AIC = Akaike information criterion; BIC = Bayesian information criterion.

* $p < .05$. ** $p < .01$.

be difficult to detect in general. It should be noted that the results from Thissen and Wainer (1982) were based on maximum likelihood estimation and the estimation employed in the current context is that of maximum a posteriori. Although Kim (2007) may provide some related information about the possible size of standard errors of c , the prior of c in Kim (2007) had the form of a beta distribution instead of the normal distribution imposed on the logit of c .

Because $G^2(1) \geq 2$ supports the augmented model by AIC, two cases for the c -difference of .10 and .15 supported the augmented model by AIC for Items 22 to 24. Because $G^2(1) \geq 7.824$ supports the augmented model by BIC, two cases for the c -difference of .10 and .15 supported the augmented model by BIC for Items 22 to 24.

Four-item anchor 3PL data. Table 6 shows the results of IRT-LR DIF analysis and information criteria with AIC and BIC from the four-item anchor 3PL data. The four items of the anchor had sets of parameters in the order of $\{a, b, c\}$ as $\{1.02, -0.84, .20\}$, $\{1.08, -0.34, .20\}$, $\{1.07, 0.40, .20\}$, and $\{1.19, 0.92, .20\}$. They had a s of about 1, approximately equality spaced b s from -0.84 to 0.92 , and c s of 0.20 . Each candidate item for which either the true b -difference of 0.30 or the true c -difference of 0.15 was analyzed for the detection of DIF. Table 6 shows all the b -difference of 0.30 were detected at $\alpha = .05$ and four of the six at $\alpha = .01$ with the IRT-LR $G^2(1)$. It shows, however, that the c -difference of 0.15 was not detected at $\alpha = .05$.

Because $G^2(1) \geq 2$ supports the augmented model by AIC, all seven cases for the b -difference of 0.30 and the c -difference of 0.15 supported the augmented model by AIC. Because $G^2(1) \geq 7.824$ supports the augmented model by BIC, two cases of candidate Items 12 and 15 for the b -difference of 0.30 supported the augmented model by BIC.

Information criteria from simulated data sets. When the $G^2(1)$ is available, the augmented model is preferred if $G^2(1) \geq 2$ by AIC, $G^2(1) \geq \log N$ by BIC, $G^2(1) \geq \log N + 1$ by CAIC, $G^2(1) \geq \log((N+2)/24)$ by RIC, $G^2(1) \geq 2N(N-1)/((N-k-2)(N-k-1))$ by AIC_c, and $G^2(1) \geq 2$ by TIC. For the current data sets, the augmented model is preferred if $G^2(1) \geq 2$ by

Table 7. Models Compared, the Likelihood Ratio Statistic, and the Model Selected by AIC and BIC.

Models compared	Likelihood ratio		Information criteria	
	G^2	df	AIC	BIC
3PL - 2PL	69.5**	12	3PL	2PL
3PL - IPL	183.8**	23	3PL	IPL
3PL - 3PL-C	9.8	11	3PL-C	3PL-C
3PL-C - 2PL	59.7**	1	3PL-C	3PL-C
3PL-C - IPL	174.0**	12	3PL-C	IPL
2PL - IPL	114.3**	11	2PL	IPL

Note. $\log N = \log 2844 = 15.90593358$. AIC = Akaike information criterion; BIC = Bayesian information criterion; 3PL = three-parameter logistic; 2PL = two-parameter logistic; IPL = one-parameter logistic. 3PL-C = 3PL with a common lower asymptote.

* $p < .05$. ** $p < .01$.

AIC, $G^2(1) \geq 7.824$ by BIC, $G^2(1) \geq 8.824$ by CAIC, $G^2(1) \geq 4.647$ by RIC, $G^2(1) \geq 2$ by AIC_c, and $G^2(1) \geq 2$ by TIC.

AIC, AIC_c, and TIC yielded consistent model selection results for data reported in Tables 3 to 6. BIC and CAIC yielded consistent model selection results for data reported in Tables 3 to 6. RIC yielded consistent model selection results with AIC for data reported in Table 4 but somewhat inconsistent model selection results with AIC for data reported in Tables 3, 5, and 6.

Results From Example Data

Model selection. The two-group example data contained in G341-19.DAT from Thissen (2001) were calibrated with 1PL, 2PL, 3PL, and 3PL-C models using MULTILOG (Thissen et al., 2002). The likelihood values reported in MULTILOG could not be used directly because the data file contained individual responses instead of the patterned data with frequencies. MULTILOG calibration runs for the 3PL and 3PL-C models used a normal prior on the $\text{logit}(c)$ with a mean of -1.1 and a standard deviation of 1 . The LR statistics G^2 could be obtained nevertheless from the outputs of the calibration runs.

Because there were 12 items, $G^2 = 69.5$ that reflected the difference between 3PL and 2PL models had 12 degrees of freedom (see Table 7). Note that the 1PL model from the marginal maximum likelihood estimation had one slope parameter and the 3PL-C model had one lower asymptote parameter to estimate. In a sense, the models presented in Table 7 are all compact models for the respective models in the context of the detection of DIF because there were two groups, focal and reference.

It should be noted that the IRT-LR statistics G^2 in the cases reported in Table 7 may not be theoretically distributed as chi-square variates with the respective degrees of freedom due to the marginalization of the seemingly different distributions. The significance testing results from G^2 , however, are presented in Table 7. When AIC was applied to these models, the models preferred were consistent with those from G^2 . Per one degree of freedom, AIC required the increment of 2 and BIC required the increment of $\log 2844 = 15.9$. Based on the G^2 and AIC results, 3PL-C might be the best model to apply to the data. BIC always selected the model with less parameters except for the comparison of the 3PL and 3PL-C models. BIC yielded the 1PL model to be the best for the data.

Table 8. 3PL-C IRT LR-DIF Analysis Results for G34I-19.DAT.

Item	Hypothesis	G^2	df	Reference group			Focal group			Focal group		
				a	b	c	a	b	c	$\hat{\mu}$	$\hat{\sigma}$	
1	All equal	2.3	2	0.87	0.49	.23	1.11	0.43	.23	0.41	1.00	
2	All equal	8.9*	2	0.93	-1.71	.23	0.60	-2.05	0.23	.43	1.00	
3	All equal	0.6	2	0.80	0.89	.23	0.91	0.83	0.23	.41	1.00	
4	All equal	24.6**	2	0.70	0.01	.23	0.78	-0.42	0.23	.38	1.00	
5	All equal	1.7	2	0.81	0.40	.23	0.67	0.49	0.23	.42	1.00	
6	All equal	0.6	2	0.82	-1.06	.23	0.91	-0.94	0.23	.42	1.00	
7	All equal	0.7	2	1.13	-1.89	.23	1.22	-1.91	0.23	.41	1.00	
8	All equal	11.9*	2	0.78	-0.69	.23	0.46	-1.38	0.23	.41	1.00	
9	All equal	5.7	2	1.17	-1.44	.23	1.31	-1.15	0.23	.44	1.00	
10	All equal	1.4	2	0.88	-1.29	.23		0.77	-1.32	.23	0.42	1.00
11	All equal	2.4	2	1.08	0.93	.23		1.06	1.09	.23	0.43	1.00
12	All equal	0.8	2	0.99	0.55	.23		0.94	0.64	.23	0.43	1.00

Note. 3PL-C = three-parameter logistic with a common lower asymptote; IRT = item response theory; IRT-LR = IRT likelihood ratio; DIF = differential item functioning.

* $p < .05$. ** $p < .01$.

DIF results. Table 8 presents results from IRT-LR DIF from the 3PL-C model from MULTILOG runs of the one compact model and 12 augmented models. In the augmented model calibration, other items except for the item under the consideration were contained in the anchor set. The LR test results each with the two degrees of freedom identified Items 2, 4, and 8 to be DIF items. The critical value of the chi-square with $\alpha = .05$ is 5.99. AIC identified Items 2, 4, 8, and 9 to be DIF items. BIC preferred the compact model and yielded no DIF items.

Table 9 presents results from IRT-LR DIF from the 3PL model from MULTILOG runs of the one compact model and 12 augmented models. Note that Thissen (2001) presented more refined and detailed DIF analysis results. Again, in the augmented model calibration, other items except for the item under the consideration were contained in the anchor set. The LR test results each with the three degrees of freedom identified Items 2, 4, and 8 to be DIF items. The critical value of the chi-square with $\alpha = .05$ is 7.81. AIC identified the same Items 2, 4, and 8 to be DIF items. BIC preferred the compact model and yielded no DIF items.

Results From Cohen et al.'s Data

Type I error and augmented model selection. Table 10 shows the number of significant LR statistics for each item at $\alpha = .05$ under the LR column. It also contains the number of the augmented model selected by AIC and BIC. Note that these numbers reflect false positive error ($\equiv$ type I error) of choosing the augmented model because the data were generated without any differences. For Item 1 under 3PL, there were three values of the significant G^2 for $N = 500$ and four values of the significant G^2 for $N = 2,000$. AIC resulted in nine cases of the augmented model selection for $N = 500$ and 10 cases for $N = 2,000$. BIC resulted in one case of the augmented model selection for $N = 500$ and none for $N = 2,000$.

The bottom of Table 10 shows that the total number of significant LR test statistics under 3PL is 175 for $N = 500$ and 164 for $N = 2,000$. As indicated in Cohen et al. (1996), the total number of the LR test performed across all the items were 2,495 for $N = 500$ and 2,484 for $N = 2,000$ because five cases and 16 cases for the respective sample sizes were failed to

Table 9. 3PL IRT LR-DIF Analysis Results for G34I-19.DAT.

Item	Hypothesis	G^2	df	Reference group			Focal group			Focal group	
				a	b	c	a	b	c	$\hat{\mu}$	$\hat{\sigma}$
1	All equal	1.7	3	0.98	0.56	.26	1.14	0.46	.24	0.40	1.00
2	All equal	10.0*	3	0.91	-1.83	.14	0.61	-1.98	.27	0.43	1.00
3	All equal	0.6	3	0.79	0.88	.22	0.97	0.88	.25	0.41	1.00
4	All equal	24.5**	3	0.74	0.07	.25	0.78	-0.43	.23	0.38	1.00
5	All equal	2.4	3	0.92	0.50	.27	0.68	0.52	.24	0.43	1.00
6	All equal	0.6	3	0.78	-1.20	.16	0.91	-0.95	.22	0.42	1.00
7	All equal	0.4	3	1.11	-1.97	.16	1.21	-1.94	.19	0.41	1.00
8	All equal	11.9*	3	0.82	-0.61	.26	0.45	-1.43	.21	0.41	1.00
9	All equal	5.4	3	1.12	-1.59	.11	1.28	-1.21	.18	0.44	1.00
10	All equal	0.9	3	0.86	-1.36	.19	0.75	-1.44	.15	0.42	1.00
11	All equal	3.9	3	0.91	0.87	.19	1.25	1.15	.26	0.43	1.00
12	All equal	1.3	3	0.88	0.47	.19	1.01	0.69	.25	0.43	1.00

Note. 3PL = three-parameter logistic; IRT = item response theory; IRT-LR = IRT likelihood ratio; DIF = differential item functioning.

* $p < .05$. ** $p < .01$.

converge to yield finite log likelihood values. Because of such facts, the total type I error rates were $175/2495 = .0701$ for $N = 500$ and $N = 2,000$ for $N = 2,000$. Also as indicated in Cohen et al., these two type I error rates were slightly higher than the theoretically expected value of .05. The false positive error rates for AIC were much higher than those for LR. The two rates for AIC are $349/2495 = .1399$ for $N = 500$ and $325/2484 = .1308$ for $164/2484 = .0660$. The false positive error rates for BIC were extremely smaller than those for LR. The rates for BIC are $62/2495 = .0248$ and $4/2484 = .0016$. The augmented model were more preferred by AIC but not really chosen by BIC.

Error rates for information criteria. When the compact model and augmented model were compared with the information criteria, the degrees of freedom was three. In terms of $G^2(3)$, the augmented model is preferred if $G^2(3) > 6$ by AIC, $G^2(3) \geq 3 \log N$ by BIC, $G^2(3) \geq 3(\log N + 1)$ by CAIC, $G^2(3) \geq 3(\log((N + 2)/24))$ by RIC, and $G^2(3) \geq 3 \times 2N(N - 1) / ((N - k - 2)(N - k - 1))$ by AIC_c. Note that TIC has the same cutoff as AIC for in the current study. The values of $G^2(3)$ for the respective information criteria are reported in Table 11.

Assuming that $G^2(3)$ is theoretically distributed as a chi-square random variable with the three degrees of freedom, the theoretical proportions of the chi-square distribution with the three degrees of freedom greater than the cutoff points of the respective information criteria are also reported in Table 11. The most pronounced pattern in Table 11 with regard to the theoretical proportions can be the high type I error rates for AIC and AIC_c and the extremely low type I error rates for BIC and RIC.

The general pattern in Table 11 for the number of cases that the augmented model was selected by information criteria is consistent for both $N = 500$ and $N = 2,000$. Again AIC and AIC_c are more likely to select the augmented model than did BIC and RIC. Note that the data were simulated as without any differential functioning over all items. Hence, all proportions in Table 11 are essentially type I errors. The lower the errors the better of the detection methods. Note that neither type II error nor the statistical power was considered in the current study.

Table 10. Generating Item Parameters (a , b , and c) for 3PL, Number of Significant LR Statistic at $\alpha = .05$, and the Numbers of Augmented Model Selected With AIC and BIC for Two Combined Sample Sizes ($N = 500$ and $2,000$).

Item	a	b	c	$N = 500$			$N = 2,000$		
				LR	AIC	BIC	LR	AIC	BIC
1	1.1	-0.7	.20	3	9	1	4	10	0
2	0.7	-0.6	.20	4	10	1	3	8	0
3	1.4	0.1	.20	4	9	1	3	4	0
4	0.9	0.9	.16	2	7	0	7	9	0
5	1.2	0.7	.12	4	11	0	3	6	0
6	1.6	1.1	.06	0	5	0	6	13	0
7	1.6	1.1	.06	4	9	1	2	9	0
8	1.6	-0.1	.16	4	7	2	5	8	1
9	1.2	0.5	.20	1	4	1	3	6	0
10	2.0	1.6	.16	5	9	1	4	5	0
11	1.0	1.6	.13	3	4	1	3	9	0
12	1.5	1.7	.09	3	5	3	4	8	0
13	1.0	0.7	.15	3	5	2	4	6	0
14	1.1	2.0	.06	4	5	1	4	5	0
15	1.1	2.4	.09	1	4	1	2	7	0
16	2.0	1.4	.11	3	4	2	6	9	0
17	1.7	1.3	.17	5	10	3	1	6	0
18	0.5	-0.6	.20	6	11	1	1	3	0
19	0.9	1.6	.11	5	7	1	4	8	1
20	1.3	0.4	.18	1	5	0	4	7	0
21	1.1	1.2	.05	4	6	0	2	8	0
22	1.2	1.1	.05	6	12	2	2	4	0
23	1.3	0.2	.20	6	7	4	4	6	0
24	1.3	0.2	.20	3	5	0	6	9	0
25	0.5	-0.8	.20	5	11	4	4	5	0
26	0.7	0.5	.20	1	4	0	2	7	0
27	0.7	0.5	.20	4	6	2	4	7	0
28	0.4	-0.4	.20	3	6	2	2	5	0
29	0.4	-0.4	.20	2	5	1	2	3	0
30	1.2	-0.5	.20	5	8	0	8	9	0
31	0.7	1.0	.20	5	5	2	2	3	0
32	0.7	-0.2	.20	4	7	2	1	7	0
33	0.7	-0.2	.20	2	7	1	2	3	0
34	0.5	0.0	.20	4	11	1	5	16	1
35	0.9	0.5	.14	3	6	1	0	7	0
36	1.1	1.4	.04	1	6	0	3	4	0
37	1.2	-0.6	.20	1	3	0	2	5	0
38	1.2	-0.6	.20	1	6	1	3	8	0
39	0.6	-0.5	.20	5	7	1	4	6	1
40	1.6	0.3	.18	6	10	2	1	6	0
41	1.1	0.0	.20	4	9	1	3	7	0
42	1.5	2.0	.06	4	18	1	7	12	0
43	1.9	1.9	.11	2	4	1	4	5	0
44	0.9	-0.5	.20	4	7	0	1	3	0
45	0.7	-0.5	.20	2	3	1	2	4	0
46	1.4	1.6	.11	4	7	2	2	4	0
47	1.4	1.6	.11	5	9	2	4	5	0
48	1.0	1.7	.08	3	5	0	6	9	0
49	1.2	1.1	.15	4	9	2	1	5	0
50	1.2	1.1	.15	7	10	1	2	7	0
Total				175	349	62	164	325	4

Note. 3PL = three-parameter logistic; LR = likelihood ratio; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table 11. The Number and the Proportion of Significant LR Statistic at $\alpha = .05$, the Numbers and the Proportions of the Augmented Model Selected by Information Criteria, and the Theoretical $G^2(3)$ Values With the Percentage Points for 3PL.

Sample size	Statistic	LR	Information criteria				
			AIC ^a	BIC	RIC	CAIC	AIC _c
N = 500	Number	175	349	62	60	128	334
	Proportion- p	.0701	.1399	.0248	.0240	.0513	.1339
	Theoretical $G^2(3)$	7.815	6.0000	18.6438	21.6438	9.1216	6.0973
	Theoretical p	.0500	.1116	.0003	.0001	.0277	.1069
N = 2,000	Number	164	325	4	3	30	319
	Proportion- p	.0660	.1308	.0016	.0012	.0121	.1284
	Theoretical $G^2(3)$	7.815	6.000	22.8027	25.8027	13.2715	6.0180
	Theoretical p	.0500	.1116	.0001	.0001	.0041	.1107

Note. LR = likelihood ratio; 3PL = three-parameter logistic; AIC = Akaike information criterion; BIC = Bayesian information criterion; RIC = Rissanen information criterion; CAIC = consistent AIC; TIC = Takeuchi information criterion.

^aAIC = TIC with the trace term k (Shibata, 1989).

Summary and Discussion

This is partly a didactic article on the use of information criteria of model selection for the detection of DIF under IRT. To this end, it provides a brief review and a summary of various information criteria that can be employed in model selection. Because many different forms of the information criteria may be used, precise mathematical definitions of the information criteria were presented but without detailed theoretical justifications. An illustration was presented using a data set from a published article for DIF detection. A comparison with two other existing DIF detection methods was presented in the illustration. In addition, simulation study results from the two published articles were used to obtain information criteria for the detection of DIF. Uses of information criteria for more general and complicated IRT modeling from previous studies were mentioned (e.g., Cohen & Cho, 2016). Hence, this article may suggest a unified framework for all model selection in IRT (e.g., dimensionality, DIF, linking, mixture modeling, etc.).

The design of the simulation data used in both illustration and the simulation study with four different data sets was that of a power study because the hypothesis tested was always false. Consequently, the DIF test statistics and information criteria should have always rejected the null hypothesis and chosen the augmented model regardless of the size of the effects reflected in the differences of item difficulty or guessing parameters. As noted in Thissen et al. (1988), the sample sizes of 500 for the focal group and 2,000 for the reference group may not be large enough considering the usual standard errors or posterior standard deviations (e.g., Kim, 2007; Thissen & Wainer, 1982). There are studies that investigated the null distributions of the IRT-LR test statistics for the DIF detection (Cohen et al., 1996; Kim & Cohen, 1998). Such papers may provide some indirect insight regarding use of information criteria for DIF detection. More study is needed before employing information criteria for the detection of DIF in routine item analysis in operational settings. In addition, the quality and the number of anchor items may be closely related to the performance of DIF detection methods (Meade & Wright, 2012). The IRT-LR method and use of information criteria for the detection of DIF can be performed in practice with iterative linking for which the detected DIF items are to be removed from the anchor.

The portion of this study that used the portions of Cohen et al.'s (1996) data mainly investigated type I and false positive error rates of the IRT-LR test and the information criteria, respectively. As reported in their study, type I error rates for 3PL were somewhat inflated. AIC clearly yielded more false positive error rates, whereas BIC yielded much smaller false positive error rates. The false positive error rates of AIC seemed to be too high to be used for the detection of DIF. It is certainly safe to say that AIC tends to prefer a more complicated model and BIC tends to choose a simpler model. There is also a concern for the power of the DIF indices. In this study, the extent to which it provides control of type II errors was not fully addressed in the context of the large scale simulation study. Power studies of the information criteria are also timely and needed to understand the full scope of information criteria for both model selection and the detection of DIF.

By considering the respective sizes of the penalty used in the information criteria when the degree of freedom is one, the propensity of which model will be supported by information criteria becomes somewhat clearer. For an ideal case in which all regularity conditions of the maximum likelihood estimation are satisfied, the values of the negative two log likelihood follow chi-square distributions. The increment to prefer the augmented model should be greater than 2 for AIC and that for BIC is 7.824 for the sample size of 2,500. These two values correspond approximately with .1572992 and .005155696 for the observed levels of significance for the chi-square distribution with one degree of freedom. Adding more parameters will be more easily supported by AIC and less likely by BIC.

It should be noted that the model selection using information criteria may not require the computation of the log likelihood functions if the IRT-LR $G^2(df)$ values are conveniently available. There is an excellent computer program IRTLRDIF (Thissen, 2001) that can produce the IRT-LR test statistic, $G^2(df)$, and the model selection based on information criteria can be indirectly performed via IRTLRDIF. Also note that the goodness of fit LR statistic called the deviance is available from several IRT computer programs including MULTILOG (Thissen et al., 2002), IRTPRO (Cai et al., 2011), and flexMIRT (Cai, 2012). These computer programs and some computational routines can be used to obtain the actual values of the log likelihood function as well as the values of information criteria.

Because issues of how to use information criteria due to the recent expansion of the IRT field to more complex and complicated modeling has yet to be fully addressed, a unified framework is clearly required. IRT model selection via information criteria is definitely an important consideration for both large scale assessment and small classroom assessment. The current investigation for the use of information criteria may provide measurement specialists and test developers with guidelines for performing detection of DIF under IRT in a theoretically consistent way with other model selection situations.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáke (Eds.), *2nd International Symposium on Information Theory* (pp. 267-281). Budapest, Hungary: Akadémiai Kiadó.

- Akaike, H. (1976). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Applications of statistics: Proceedings of the symposium held at Wright State University, Dayton, Ohio*. (pp. 27-53) Amsterdam, The Netherlands: North-Holland.
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.
- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: The MIT Press.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459; 47, 369 (Errata).
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.
- Bock, R. D., & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 469-513). Amsterdam, The Netherlands: Elsevier.
- Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.
- Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information-theoretic approach*. New York, NY: Springer-Verlag.
- Cai, L. (2012). flexMIRT: A numerical engine for multilevel item factor analysis and test scoring (Version 1.88) [Computer software]. Seattle, WA: Vector Psychometric Group.
- Cai, L., Thissen, D., & du Toit, M. (2011). IRTPRO: Item response theory for patient-reported outcomes [Computer software]. Lincolnwood, IL: Scientific Software International.
- Cohen, A. S., & Cho, S.-J. (2016). Information criteria. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 2, pp. 363-378). Boca Raton, FL: CRC Press.
- 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.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.
- deLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics, Volume I: Foundations and basic theory* (pp. 599-609). New York, NY: Springer-Verlag.
- Fujikoshi, Y., & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, 84, 707-716.
- Holland, P. W., & H., Wainer (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.
- Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics* (2nd ed.). New York, NY: John Wiley.
- Kang, T.-H., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, 31, 331-358.
- Kang, T.-H., Cohen, A. S., & Sung, H.-J. (2009). IRT model selection methods for polytomous items. *Applied Psychological Measurement*, 33, 499-518.
- Kim, S.-H. (2007). Some posterior standard deviations in item response theory. *Educational and Psychological Measurement*, 67, 258-279.
- Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, 22, 345-355.
- Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74, 21-48.
- Kullback, S. (1959). *Information theory and statistics*. New York, NY: John Wiley.
- Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement*, 33, 353-373.
- Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020.

- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, 40, 111-135.
- May, H. (2006). A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics*, 31, 63-79.
- Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, 97, 1016-1031.
- Parzen, E., Tanabe, K., & Kitagawa, G. (Eds.). (1998). *Selected papers of Hirotugu Akaike*. New York, NY: Springer-Verlag.
- Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.
- Rao, C. R., & Wu, Y. (2001). On model selection (with discussions and rejoinder). In P. Lahiri (Ed.), *Model selection* (pp. 1-64). Beachwood, OH: Institute of Mathematical Statistics.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465-471.
- Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. Tokyo, Japan: KTK Scientific Publishers.
- Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343.
- Shibata, R. (1989). *Statistical aspects of model selection* (IIASA Working Paper, No. WP-89-077). Laxenburg, Austria: International Institute for Applied Systems Analysis. Retrieved from <http://pure.iiasa.ac.at/3267/1/WP-89-077.pdf>
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64, 353-616.
- Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics: Part A-theory and Methods*, A, 7, 13-26.
- Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175-186.
- Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory* (Version 6.0). Chicago, IL: Scientific Software.
- Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina.
- Thissen, D., Chen, W.-H., & Bock, R. D. (2002). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* [Computer software]. Lincolnwood, IL: Scientific Software International.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group differences: The concept of item bias. *Psychological Bulletin*, 99, 118-128.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.
- Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, 47, 397-412.
- Tutz, G., & Schauberer, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, 80, 21-43.
- Venables, W. N., Smith, D. M., & The R Development Core Team. (2009). *An introduction to R* (2nd ed.). La Vergne, TN: Network Theory.