Psychological Test and Assessment Modeling, Volume 59, 2017 (2), 183-205

Performances of LOO and WAIC as IRT model selection methods

Yong Luo[1] *& Khaleel Al-Harbi*[2,3]

Abstract

The widely available information criterion (WAIC) and leave-one-out cross-validation (LOO) are considered fully Bayesian model selection methods due to their utilization of the whole posterior distribution other than the point estimates. Despite their theoretical advantage of being fully Bayesian, how such an advantage translates into practical performance remains unknown. In this paper, we conducted a simulation study to compare the performances of WAIC and LOO with other four commonly used methods, which are the likelihood ratio test (LRT), AIC, BIC, and DIC, in the context of dichotomous IRT model selection. We also used a real data set to illustrate that those six model selection methods can lead to different conclusions. The findings suggest that WAIC and LOO perform better than the other four methods, especially when the data were generated with 3PLM. In addition, it was found that AIC, one of the most widely used model selection method, can become inconsistent with different sample sizes and test lengths.

Key words: information criterion, model selection, IRT, WAIC, LOO

[1] *Correspondence concerning this article should be addressed to:* Yong Luo, PhD, National Center for Assessment - West Palm neighborhood j - King Khalid bin Abdul Aziz P.O. Box 68566, Riyadh 11534, Saudi Arabia; email: jackyluoyong@gmail.com

[2] National Center for Assessment, Riyadh, Saudi Arabia

[3] Taibah University, Medina, Saudi Arabia

Introduction

The field of Bayesian item response theory (IRT) has been growing rapidly since Albert's seminal paper (1992), in which he showed that the two-parameter normal ogive model can be estimated with the Gibbs sampling, and Patz and Junker's groundbreaking papers (1999a, 1999b), in which they proposed Metroplis-Hastings algorithm for estimation of various dichotomous and polytomous IRT models. The development of those Monte Carlo Markov Chain (MCMC) methods and their implementation in various software programs have made Bayesian IRT increasingly popular among researchers and practitioners.

In contrast to the frequentist IRT paradigm in which point estimates of item and ability parameters are of interest, with Bayesian IRT we obtain posterior distributions of model parameters, upon which further analysis can be conducted. Sometimes posterior distributions are used to gain point summarizations such as Expected A Posteriori (EAP) and Maximum A Posteriori (MAP), which can be viewed as the Bayesian analogs of the point estimate of a model parameter obtained through marginal maximum likelihood estimation (MMLE) with EM algorithm (Bock, & Aitkin, 1981) in the frequentist IRT paradigm. It should be noted that, however, EAP and MAP are merely point summarization of the corresponding posterior distribution and are not, in any regard, intended to be estimates of a true model parameter that is believed to exist in the frequentist paradigm.

With an expanding body of literature on Bayesian IRT (e.g., Beguin & Glas, 2001; Bolt & Lall, 2003; Cao & Stokes, 2008; Fox & Glas, 2003; Jiao & Zhang, 2015), the topic of Bayesian model checks and comparison in the IRT context has also received increasing attention (e.g., Levy, Mislevy, & Sinharay, 2009; Li, Xie, & Jiao, 2016; Sinharay, 2005; Sinharay, Johnson, & Stern, 2006; Zhu & Stone, 2011). As nicely summarized by Sinharay (2016), some popular Bayesian model checking techniques that have been applied to IRT include Bayesian residual analysis, prior predictive checks, and posterior predictive checks, and Bayesian model comparison methods include the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), the Bayes factor (Kass & Raftery, 1995), and cross-validation likelihood and partial Bayes factor (O'Hagan, 1995). Complementary in nature, Bayesian model checks and comparison serve different purposes: the former allow a researcher to examine whether a model captures the important features of a data set, while the latter answers the question of which model out of a group of candidate models fits a data set the best. In this paper, we focus on Bayesian model comparison methods.

In the Bayesian IRT literature, one of the most popular model comparison methods is DIC. DIC is more Bayesian than other commonly used model comparison methods such as Akaike's information criterion (AIC; Akaike, 1973, 1974) and Bayesian information criterion (BIC; Shwarz, 1978) in that its computation requires use of the whole posterior distribution. However, as will be discussed later, DIC computes its deviance term based on EAP other than the whole posterior distribution and hence should only be considered partially Bayesian. A truly fully Bayesian model selection method does not involve point estimates in its computation. The widely available information criterion (WAIC; Watanabe, 2010) is such a fully Bayesian model selection method. As its name suggests,

WAIC is based on information criterion and is considered an improved version of DIC (Vehtari, Gelman, & Gabry, 2016a) because "WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate" (Gelman, Hwang, & Vehtari, 2013, p. 1003). WAIC utilizes the whole posterior distributions to compute its deviance and penalty terms (which will be shown later), and hence is believed to be superior to other information criterion based model selection methods that use point estimates in the computation.

Another fully Bayesian model comparison method is the Bayesian leave-one-out cross-validation (LOO; Geiser & Eddy, 1979), to which AIC, DIC, and WAIC have been shown to be asymptotically equal (Shibata, 1989; Stone, 1977; Watanabe, 2010). LOO requires taking out one data point at a time and using it to cross-validate the model estimated with the remaining data. Due to its iterative nature, LOO can be computationally prohibitive for large sample datasets in that the model needs to be fitted for n (the sample size) times and computational shortcuts such as importance sampling techniques (e.g., Gelfand, Day, & Chang, 1992; Ionides, 2008) have been proposed to approximate the posterior distribution without re-fitting the model for multiple times. In this paper we evaluate the performance of LOO computed with a relatively new importance sampling technique, namely Pareto smoothed importance sampling (PSLS; Vehtari, Gelman, & Gabry, 2015)[4].

Numerous model selection methods have been used in the context of IRT model comparison and selection, and the performances of some of those methods in selecting the correct IRT model have also been systematically investigated using simulation studies. To date, WAIC and LOO have not been applied to choosing IRT models in empirical studies, nor have their performances as model selection methods been studied in simulation studies. This study aims to fill the gap in the psychometric literature by investigating how WAIC and LOO perform, in comparison with other four commonly used methods, in an IRT model selection scenario through a simulation study. We intend to address two research questions in this study. First, how does the theoretical superiority of WAIC and LOO translate into practical performance in the context of dichotomous IRT model comparison and selection? Second, does WAIC perform equally well as LOO in our simulated conditions?

The remainder of this article is organized into five sections. We start with a literature review of IRT related simulation studies that investigate the performances of various model selection methods. Then we provide a brief description of the six methods used in this study. In the third section, we conduct a simulation study to compare the performances of WAIC and LOO with the other four methods, followed by a demonstration using a real data set in the fourth section. We close our paper with conclusions and discussions of the implications of using WAIC and LOO in practice.

[4] Due to space limitation and its technical nature, we do not describe the PSLS and its implementation here. Interested readers are referred to Vehtari, Gelman, and Gabry (2015, 2016a).

Model selection studies in IRT literature

Model selection methods, especially those based on information criterion, have been widely applied in the IRT literature. As nicely summarized by Cohen and Cho (2016), applications of information criterion based methods to IRT model comparison and selection fall into the following five categories: exploring the number of dimensions (e.g., Yao & Schwarz, 2006), a general model with different constraints (e.g., Hickendorff, Heiser, van Putten, & Verhelst, 2009), nested models (e.g., Revuelta, 2008), multilevel IRT models (e.g., May, 2006), and IRT models from different families (e.g., Rijmen & De Boeck, 2002).

While there are a large number of studies applying model selection methods to IRT model selection, only a limited number of them (Kang & Cohen, 2007; Kang, Cohen, & Sung, 2009; Li, Bolt, & Fu, 2006; Li, Cohen, Kim, & Cho, 2009; Whittaker, Chang, & Dodd, 2012, 2013; Zhu, & Stone, 2012) in the literature have systematically investigated the performances of different model selection methods with simulation studies.

Kang and Cohen (2007) conducted a simulation study to compare the performances of the likelihood ratio test (LRT), AIC, BIC, DIC, and the cross-validation log-likelihood (CVLL; O'Hagan, 1995) in selecting the correct model among the one-parameter logistic model (1PLM), the two-parameter logistic model (2PLM), and the three-parameter logistic model (3PLM). They found that those five methods tend to produce inconsistent results and overall, CVLL is the best model selection method in their simulation conditions.

In another simulation study, Kang, Cohen, and Sung (2009) compared the performances of AIC, BIC, DIC, and CVLL in selecting the correct model among the graded response model (GRM; Samejima, 1969), the partial credit model (PCM; Masters, 1982), the generalized partial credit model (GPCM; Muraki, 1992), and the rating scale model (RSM; Andrich, 1978). They found that while the simulated condition affects how well each method performs, CVLL, on average, has the best performance.

Whittaker, Chang, and Dodd (2012) investigated the performances of LRT, AIC, the finite sample corrected AIC (AICC; Hurvich & Tsai, 1989), BIC, Hannon and Quinn's information criterion (HQIC; Hannon & Quinn, 1979), and consistent AIC (CAIC; Bozdogan, 1987) in selecting the correct model combination among several competing mixed-format IRT models in a simulation study. It was found that those methods generally perform well in their simulation conditions, although they tend to choose less parameterized models when the generating model is more parameterized. Whittaker, Chang, and Dodd (2013) replicated their previous simulation study by increasing the variance of the generating item discrimination parameters and found that those methods perform well in all simulation conditions. They concluded that the anomaly found in their previous study is due to the small variance of the generating item discrimination parameters.

Whereas the previous four studies focus on the performances of different methods in selecting a correct unidimensional IRT model, the following studies target more complex IRT models. Li, Cohen, Kim, and Cho (2009) investigated the performances of AIC,

BIC, DIC, the pseudo-Bayes Factor (PsBF; Gerisser & Eddy, 1979; Gelfand, Dey, & Chang, 1992), and posterior predictive model checks (PPMC; Gelman, Meng, & Stern, 1996) in selecting the correct model among several competing mixture IRT models. They found that BIC and PsBF are most effective; AIC and PPMC tend to choose more complex models in some simulating conditions; DIC is the least effective method. Li, Bolt and Fu (2006) compared the performances of PsBF, DIC, and posterior predictive checks (Gelman, Carlin, Stern, & Rubin, 2014) in selecting the correct model among different testlet models, and found that DIC performs worse than the other two methods. Zhu and Stone (2012) investigated the performances of DIC, conditional predictive ordinate (CPO), and PPMC in selecting among GRM and several alternative models that are modified or extended from GRM, which include the one-parameter GRM (Muraki, 1990), two-dimensional GRM with simple- and complex structures, and the testlet version of the GR model. They found that the three methods perform equally well in their simulated conditions.

To date, there have been no studies in the IRT literature that investigate performances of WAIC and LOO in the context of IRT model comparison and selection. In this study, we focus on dichotomous IRT models and compare the performances of WAIC and LOO with other four commonly used methods, namely LRT, AIC, BIC, and DIC. In the next section, we provide a brief description of the six model selection methods.

Model selection methods in the current study

The six model selection methods that are of interest in this study fall into either the frequentist or the Bayesian framework. Among them, LRT, AIC, and BIC are frequentist, and DIC, WAIC, and LOO are Bayesian. LRT is based on a deviance term that can be calculated using the following equation

$$Deviance = -2\log p(y \mid \hat{\theta}_{mle}), \tag{1}$$

where $\hat{\theta}_{mle}$ is the maximum likelihood estimate, and $\log p(y \mid \hat{\theta}_{mle})$ is the log likelihood based on the maximum likelihood estimate $\hat{\theta}_{mle}$. LRT only applies to model selection among a group of nested models. The reason is that the deviance difference between two nested models follows a chi-square distribution, and therefore can be used as a test statistic to test whether the more parameterized model fits the data significantly better than the less parameterized one. Considering that there are nested relations among the three common dichotomous IRT models, LRT is applicable and therefore included in the current study.

As one of the most widely used information-criterion-based model selection methods, AIC is more flexible than LRT and can be applied to scenarios where competing models are not nested. The computation of AIC is given as

$$AIC = -2\log p(y \mid \hat{\theta}_{mle}) + 2k, \tag{2}$$

where $-2\log p(y|\hat{\theta}_{mle})$ is the same deviance term as in equation 1, and k is the number of estimated parameters with $2k$ serving as a penalty term for model complexity.

Another common information-criterion-based model selection method is BIC. The computation of BIC is given as

$$BIC = -2\log p(y|\hat{\theta}_{mle}) + k(\log N), \tag{3}$$

where N is the model sample size and the other terms remain the same as in equation 2. AIC and BIC only differ in the penalty term: for BIC the penalty term increases with the increase of sample size; for AIC, it remains constant regardless of the sample size. The penalty terms of AIC and BIC are equal when the sample size is 100 (log100 = 2), and when the sample size is larger than 100, BIC imposes a harsher penalty for model complexity.

DIC can be computed using the following equation

$$DIC = -2\log p(y|\hat{\theta}_{EAP}) + 2p_{DIC}, \tag{4}$$

where $-2\log p(y|\hat{\theta}_{EAP})$ is the deviance term based on the posterior mean estimate $\hat{\theta}_{EAP}$ and the computation of p_{DIC} is given as

$$p_{DIC} = 2(\log p(y|\hat{\theta}_{EAP}) - E_{post}(\log p(y|\theta))). \tag{5}$$

In equation 5, $E_{post}(\log p(y|\theta))$ is the posterior mean of the log likelihood and can be computed with the following equation

$$E_{post}(\log p(y|\theta)) = \frac{1}{S}\sum_{s=1}^{S}\log p(y|\theta^s), \tag{6}$$

where S is the number of simulation draws and θ^s is the simulated value for parameter θ at the sth draw. As mentioned in the introduction, DIC is only partially Bayesian in that the computation of its deviance term in equation 4 and the first term of its penalty term in equation 5 is based on $\hat{\theta}_{EAP}$, which is the point summarization of the posterior distribution, and only the computation of the second term of its penalty term (equation 6) utilizes the whole posterior distribution.

Before we discuss the computation of LOO and WAIC, it is necessary to introduce log pointwise predictive density (LPPD), which can be computed with the following equation (Gelman et al., 2014)

$$LPPD = \sum_{i=1}^{n}\log\int p(y_i|\theta)p_{post}(\theta)d\theta. \tag{7}$$

LPPD can be viewed as a fully Bayesian analog of the term $\log p(y|\hat{\theta}_{mle})$ used in the computation of AIC and BIC, or the term $\log p(y|\hat{\theta}_{EAP})$ in the computation of DIC. It is

fully Bayesian because of its utilization of $p_{post}(\theta)$, which is the posterior distribution of parameters, other than the use of point estimates $\hat{\theta}_{mle}$ or $\hat{\theta}_{EAP}$. According to Gelman et al., LPPD can be computed as

$$LPPD_{computed} = \sum_{i=1}^{n} \log(\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta)) . \quad (8)$$

With LPPD defined, the computation of WAIC is given as

$$WAIC = -2LPPD + 2p_{WAIC} , \quad (9)$$

where p_{WAIC} is the penalty term and can be computed in the following two ways:

$$p_{WAIC1} = 2\sum_{i=1}^{n} (\log(E_{post} p(y_i \mid \theta)) - E_{post}(\log p(y_i \mid \theta))) , \quad (10)$$

$$p_{WAIC2} = \sum_{i=1}^{n} \mathrm{var}_{post}(\log p(y_i \mid \theta)) . \quad (11)$$

According to Gelman et al., the second approach (equation 11) is more computationally stable and is implemented in the R package loo (Vehtari, Gelman, & Gabry, 2016b), which will be discussed later. As can be seen from equations 7-11, the computation of both LPPD and the penalty term of WAIC utilizes the whole posterior distribution other than point estimates, which is why WAIC is considered fully Bayesian.

In cross validation, a dataset is partitioned into a training set and a validation set. Usually we fit a model of interest to the training set and obtain a posterior distribution, with which we evaluate the fit of the model to the validation set. A special case of cross validation is leave one out cross validation, in which we leave one data point out each time and compute LPPD with *n-1* data points using the following equation (Gelman et al., 2014)

$$LPPD_{loo} = \sum_{i=1}^{n} \log p_{post(-i)}(y_i \mid \theta) , \quad (12)$$

where *n* is the sample size, and $\log p_{post(-i)}(y_i \mid \theta)$ is the log likelihood of the *i*th dataset that excludes the *i*th data point, whose computation is given by Gelman et al. as

$$\sum_{i=1}^{n} \log p_{post(-i)}(y_i \mid \theta) = \sum_{i=1}^{n} \log(\frac{1}{S}\sum_{s=1}^{S} p(y_i \mid \theta^{is})) . \quad (13)$$

In equation 13, θ^{is} is the *s*th simulated value in the posterior distribution conditioning on the *i*th dataset without the *i*th data point. LOO is further computed as -2 times to be on the same scale as AIC, BIC, DIC, and WAIC. Similarly, the computation of LOO requires the use of the whole posterior distribution (equation 13) and therefore, LOO is also fully Bayesian.

Methods

Simulation design

The current simulation study is similar to the one conducted by Kang and Cohen (2007), with CVLL replaced by WAIC and LOO. In the following, we provide a detailed description of their simulation study. To compare the performances of LRT, AIC, BIC, DIC, and CVLL in the context of dichotomous IRT model comparison and selection, Kang and Cohen conducted a simulation study with 36 conditions (two test lengths × two sample sizes × three ability distributions × three generating dichotomous IRT models). The two test lengths included 20 and 40 items; the two sample sizes were 500 and 1000; the three ability distributions were $N(-1, 1)$, $N(0, 1)$, and $N(1, 1)$; and the three generating IRT models were 1PLM, 2PLM, and 3PLM. The item difficulty parameters were generated from a standard normal distribution $N(0, 1)$, item discrimination parameters from a lognormal distribution $lnN(0, 0.5)$, and item pseudo-guessing parameters

Table 1:
Item Parameters Used for Data Generation

Item	a	b	c	Item	a	b	c
Item 1	1.1005	0.4078	0.2228	Item 21	0.5659	-0.1257	0.3426
Item 2	2.2093	0.5696	0.2332	Item 22	0.6128	-0.7826	0.1925
Item 3	1.4493	-1.061	0.2337	Item 23	1.1037	0.0615	0.2324
Item 4	0.7514	-0.2437	0.1445	Item 24	1.9886	0.4244	0.1396
Item 5	1.5789	0.3206	0.2581	Item 25	0.5691	-0.735	0.2059
Item 6	0.6425	-1.3762	0.2712	Item 26	1.0346	0.9836	0.3124
Item 7	1.6254	-0.98	0.1232	Item 27	1.1384	-1.2651	0.1832
Item 8	1.3415	-0.6881	0.1954	Item 28	3.3488	-0.2252	0.1811
Item 9	0.918	-0.3526	0.2709	Item 29	2.6306	-0.6576	0.2537
Item 10	1.8027	0.24	0.2984	Item 30	0.6652	1.7007	0.2184
Item 11	0.8159	0.5917	0.0587	Item 31	1.0342	1.0805	0.2261
Item 12	0.9375	1.8891	0.1405	Item 32	1.0163	-2.0452	0.3464
Item 13	0.9126	-0.269	0.2339	Item 33	1.2945	0.1627	0.1455
Item 14	1.9395	0.3673	0.2387	Item 34	1.6521	0.0573	0.3861
Item 15	0.3746	-0.9681	0.3527	Item 35	0.9696	1.2171	0.1046
Item 16	0.673	-1.2601	0.1206	Item 36	1.2369	2.1226	0.1656
Item 17	0.4166	0.5225	0.1244	Item 37	0.7812	0.4228	0.2696
Item 18	1.2093	-1.3356	0.1167	Item 38	0.7728	-0.1656	0.178
Item 19	0.9486	0.9515	0.2787	Item 39	0.5441	-0.2055	0.1961
Item 20	1.4916	0.9811	0.1923	Item 40	1.4025	1.2841	0.2917

Note. a is item discrimination parameter; b is item difficulty parameter; c is pseudo-guessing parameter.

from a beta distribution *beta*(5, 17). When the generating IRT model was 1PLM, only the difficulty parameters were used with discrimination parameters all fixed to 1; when the generating IRT model was 2PLM, both the difficulty and discrimination parameters were used while the pseudo-guessing parameter was fixed to 0. When the test length was 20 items, only the first 20 item parameters were used in generating item responses. The specific item parameters they generated are listed in Table 1.

To compare the performances of those five model selection methods, Kang and Cohen ran 50 replications in each condition and computed the proportion of times when the generating model was selected as the best fitting model by each of the five methods.

In the current simulation study, we use the same simulation design with the aforementioned 36 conditions, and in each condition we generated 50 datasets using the item parameters listed in Table 1. It should be noted that we set the number of replications within each simulation condition to 50 to keep the current study at a manageable level, due to the long computation time required for the implementation of MCMC algorithm. It is not uncommon for IRT simulation studies using MCMC algorithms to run only 25 replications per condition (e.g., Jiao, Wang, & He, 2013), which is the minimum number of replications in IRT simulation studies considered acceptable (Harwell, Stone, Hsu, & Kirisci, 1996). For each generated dataset, we fit 1PLM, 2PLM, and 3PLM and compare the model fit using the six methods. Within each condition we record how many times each method chooses a model, and divide the number of times a correct model is chosen by 50 to obtain the power rate of a given method. The power rate is the dependent variable of the simulation study.

Estimation methods

For the computation of LRT, AIC, and BIC, we use the R package **mirt** (Chalmers, 2012) that implements MMLE method. OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2010) is used for the computation of DIC. The R package **rstan**, which is the R interface to Stan (Carpenter et al., 2016), is used for model estimation and the R package **loo** (Vehtari et al., 2016b) is used for the computation of WAIC and LOO. It should be noted that Stan is a relatively new statistical software program that implements the no-U-turn sampler (NUTS; Hoffman & Gelman, 2014), an extension to a powerful and efficient MCMC algorithm called Hamiltonian Monte Carlo (HMC; Neal, 2011). Luo and Jiao (2016) provided a collection of Stan code for common dichotomous and polytomous IRT models and showed that Stan is an attractive alternative for Bayesian IRT model estimation.

Estimation of the three IRT models with MCMC methods requires the specification of prior distributions for all model parameters, and we use priors similar to those used by Kang and Cohen (2007). For the 1PLM model, we assign a standard normal distribution $N(0, 1)$ as the prior for the ability parameters for model identification, and a normal distribution with unknown mean and variance as the prior for the item difficulty parameter; the unknown mean is assigned the distribution $N(0, 25)$ as the hyperprior, and for the unknown variance, in OpenBUGS we assign the distribution *gamma*(1, 1) as the hyper-

prior for the precision (reciprocal of the variance), and in Stan a half Cauchy distribution $Cauchy_{+}(0, 5)$ as the hyperprior for the standard deviation (square root of the variance). For 2PLM and 3PLM, we use same priors for parameters already in the 1PLM model, and assign the distribution truncated normal distribution $N_{+}(0, 4)$ as the prior for the item discrimination parameter, and a beta distribution $beta(5, 23)$ for the pseudo-guessing parameter. It should be noted that the above prior choices are recommended and implemented in the context of Bayesian IRT (e.g., Levy & Mislevy, 2016; Sahu, 2002; Sheng, 2010), and we choose them so that the results regarding the performances of WAIC and LOO in the current study are relevant to practitioners and researchers who are likely to use the same priors. While a topic warranting further studies, the effect of prior choices upon the performances of WAIC and LOO is out of the scope of the current study and will not be further discussed in the remainder of this paper.

Model convergence check

To check convergence for models estimated with MCMC methods, we apply the Gelman and Rubin's convergence diagnostic (Gelman & Rubin, 1992) that computes the potential scale reduction factor (PSRF). A PSRF value close to 1 indicates model convergence and in practice, the value of 1.1 has been recommended as the threshold to gauge whether the model has converged (Gelman, Carlin, Stern, & Rubin, 2014). In OpenBUGS all PSRF values converge to 1 within 2,000 iterations for the three IRT models, and we run three parallel chains with 5,000 iterations each to be conservative. The efficient HMC algorithm implemented in Stan needs approximately 200 iterations for the three models to converge, and to be conservative we run three parallel chains with 500 iterations each in Stan to ensure that model convergence is not a concern.

Results

Table 2 lists the model selection results based on the six methods. Specifically, it provides the number of times a given method chooses 1PLM, 2PLM, and 3PLM in a simulation condition. For example, the first row of Table 2 shows that when the generating model is 1PLM with a sample size of 500 (simulated from a $N(-1, 1)$ distribution) and a test length of 20, LRT chooses 1PLM 46 times and 2PLM 4 times; AIC, BIC, DIC, and LOO choose 1PLM 50 times; WAIC chooses 1PLM 49 times and 2PLM once.

Figure 1 provides a visual comparison of the mean power rates comparison of the six methods across 36 simulation conditions. As can be seen, LOO and WAIC have the highest power (power = 0.98) and the performance of DIC is slightly worse (power = 0.93); the other three methods in the frequentist framework perform considerably worse in that LRT (power = 0.88) performs slightly better than AIC (power = 0.85), and the average performance of BIC is the worst (power = 0.67).

Table 2:
Model Selection Results

Sample Size	Ability Distribution	Test Length	True Model	Model Selection Methods																	
				LRT			AIC			BIC			DIC			LOO			WAIC		
				1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
		20	1	46	4	0	50	0	0	50	0	0	50	0	0	50	0	0	49	1	0
	N(-1, 1)		2	0	50	0	0	50	0	1	49	0	0	50	0	0	50	0	0	50	0
			3	0	11	39	0	20	30	50	0	0	0	0	50	0	0	50	0	0	50
		40	1	46	4	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0
			2	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0
			3	0	0	50	0	3	47	43	7	0	0	0	50	0	0	49	0	1	49
		20	1	46	4	0	49	1	0	50	0	0	50	0	0	48	2	0	48	2	0
500	*N*(0, 1)		2	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0
			3	0	7	43	0	27	23	36	14	0	0	8	42	0	0	50	0	0	50
		40	1	47	3	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0
			2	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0
			3	0	0	50	0	21	29	0	50	0	0	0	50	0	0	50	0	0	50
		20	1	46	4	0	49	1	0	50	0	0	50	0	0	48	1	1	48	1	1
	N(1, 1)		2	0	50	0	0	50	0	0	50	0	0	50	0	0	47	3	0	45	5
			3	0	48	2	0	49	1	29	21	0	0	31	19	0	6	44	0	3	47
		40	1	47	3	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0
			2	0	50	0	0	50	0	0	50	0	0	48	2	0	50	0	0	50	0
			3	0	36	14	0	50	0	0	50	0	0	29	21	0	4	46	0	2	48
		20	1	42	8	0	50	0	0	50	0	0	49	1	0	49	1	0	45	5	0
	N(-1, 1)		2	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0
			3	0	1	49	0	2	48	14	36	0	0	1	49	0	0	50	0	0	50
		40	1	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0
			2	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0
			3	0	0	50	0	0	50	0	50	0	0	0	50	0	0	50	0	0	50
		20	1	44	6	0	50	0	0	50	0	0	50	0	0	49	1	0	47	3	0
1000	*N*(0, 1)		2	0	50	0	0	50	0	0	50	0	0	50	0	0	49	1	0	50	0
			3	0	0	50	0	2	48	0	50	0	0	2	48	0	0	50	0	0	50
		40	1	47	3	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0
			2	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0
			3	0	0	50	0	0	50	0	50	0	0	0	50	0	0	50	0	0	50
		20	1	45	5	0	49	1	0	50	0	0	50	0	0	48	2	0	46	4	0
	N(1, 1)		2	0	50	0	0	50	0	0	50	0	0	49	1	0	50	0	0	50	0
			3	0	45	5	0	49	1	0	50	0	0	45	5	0	9	41	0	5	45
		40	1	49	1	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0
			2	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0
			3	0	22	28	0	50	0	0	50	0	0	13	37	0	3	47	0	2	48

Note. The maximum selection rate within a cell is 50. In the True Model column, 1 denotes 1PLM, 2 2PLM, and 3 3PLM.

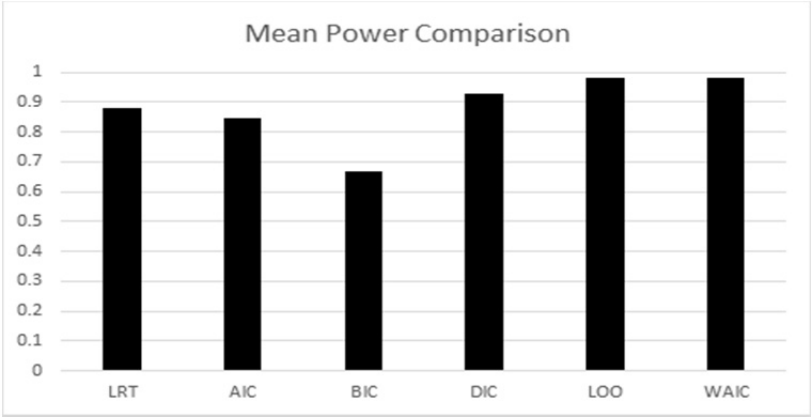

Figure 1:
Mean power rates comparison by methods

Table 3 lists the power rates of each method in different simulating conditions, and tells how the choice of a generating model affects the performance of certain methods. For example, when the generating model is 1PLM the power of BIC is constantly one; when the generating model is 3PLM, however, its power becomes invariably zero. Another noticeable finding is that when the ability distribution is $N(1, 1)$ and the generating model is 3PLM, WAIC and LOO have considerably higher power than the other four traditional methods, regardless of the sample size and test length.

In Figures 2-4, we provide a visual presentation of the marginal summary of the performances of the six methods using bar plots. Specifically, in Figure 2 we consider the mean power rates of a given test length combined with a given generating model by averaging the power rates across six simulation conditions (three ability distributions × two sample sizes). When the generating model is 1PLM with the test length of 20, LOO and WAIC perform slightly worse than AIC, BIC and DIC, and WAIC has slightly lower power than LOO; when the test length increases to 40, LOO and WAIC perform as well as AIC, BIC, and DIC, whereas LRT performs worse. When the generating model is 2PLM, all six methods perform approximately equally well. When the generating model is 3PLM, regardless of the test length both WAIC and LOO perform very well and have power rates above 0.90, while the performances of LRT, AIC, and DIC, which improve when the test length changes from 20 to 40, are considerably worse.

In Figure 3 we consider the mean power rates of the combination of a given sample size and a given generating model by averaging across six simulation conditions (three ability distributions × two test lengths). When the generating model is 1PLM, LOO and WAIC perform slightly worse than AIC, BIC and DIC, and WAIC has slightly lower power than LOO, although the difference is negligible; LRT performs worse than the other five methods. When the generating model is 2PLM, all six methods perform approximately

Table 3:
Power Rates of Different Methods

Sample Size	Ability Distribution	Test Length	True Model	Model Selection Methods					
				LRT	AIC	BIC	DIC	LOO	WAIC
500	*N*(-1, 1)	20	1	0.92	1	1	1	1	0.98
			2	1	1	0.98	1	1	1
			3	0.78	0.6	0	1	1	1
		40	1	0.92	1	1	1	1	1
			2	1	1	1	1	1	1
			3	1	0.94	0	1	0.98	0.98
	N(0, 1)	20	1	0.92	0.98	1	1	0.96	0.96
			2	1	1	1	1	1	1
			3	0.86	0.46	0	0.84	1	1
		40	1	0.94	1	1	1	1	1
			2	1	1	1	1	1	1
			3	1	0.58	0	1	1	1
	N(1, 1)	20	1	0.92	0.98	1	1	0.96	0.96
			2	1	1	1	1	0.94	0.9
			3	0.04	0.02	0	0.38	0.88	0.94
		40	1	0.94	1	1	1	1	1
			2	1	1	1	0.96	1	1
			3	0.28	0	0	0.42	0.92	0.96
1000	*N*(-1, 1)	20	1	0.84	1	1	0.98	0.98	0.9
			2	1	1	1	1	1	1
			3	0.98	0.96	0	0.98	1	1
		40	1	1	1	1	1	1	1
			2	1	1	1	1	1	1
			3	1	1	0	1	1	1
	N(0, 1)	20	1	0.88	1	1	1	0.98	0.94
			2	1	1	1	1	0.98	1
			3	1	0.96	0	0.96	1	1
		40	1	0.94	1	1	1	1	1
			2	1	1	1	1	1	1
			3	1	1	0	1	1	1
	N(1, 1)	20	1	0.9	0.98	1	1	0.96	0.92
			2	1	1	1	0.98	1	1
			3	0.1	0.02	0	0.1	0.82	0.9
		40	1	0.98	1	1	1	1	1
			2	1	1	1	1	1	1
			3	0.56	0	0	0.74	0.94	0.96

Note. In the True Model column, 1 denotes 1PLM, 2 2PLM, and 3 3PLM.

equally well. When the generating model is 3PLM, regardless of the sample size both WAIC and LOO perform very well and have power rates well above 0.90, while the performances of LRT, AIC, and DIC, which improve slightly when the sample size changes from 500 to 1000, are noticeably worse.

In Figure 4 we consider the mean power rates of the combination of a given ability distribution condition (whether it matches with item difficulty) and a given generating model by averaging across four simulation conditions (two sample sizes × 2 test lengths). Similar to Figure 3, when the generating model is 1PLM, LOO and WAIC perform slightly worse than AIC, BIC and DIC, and WAIC has slightly lower power than LOO, although the difference is negligible. When the generating model is 2PLM, all six methods perform approximately equally well. When the generating model is 3PLM, regardless of the ability distribution both WAIC and LOO perform very well and have power rates well above 0.90,

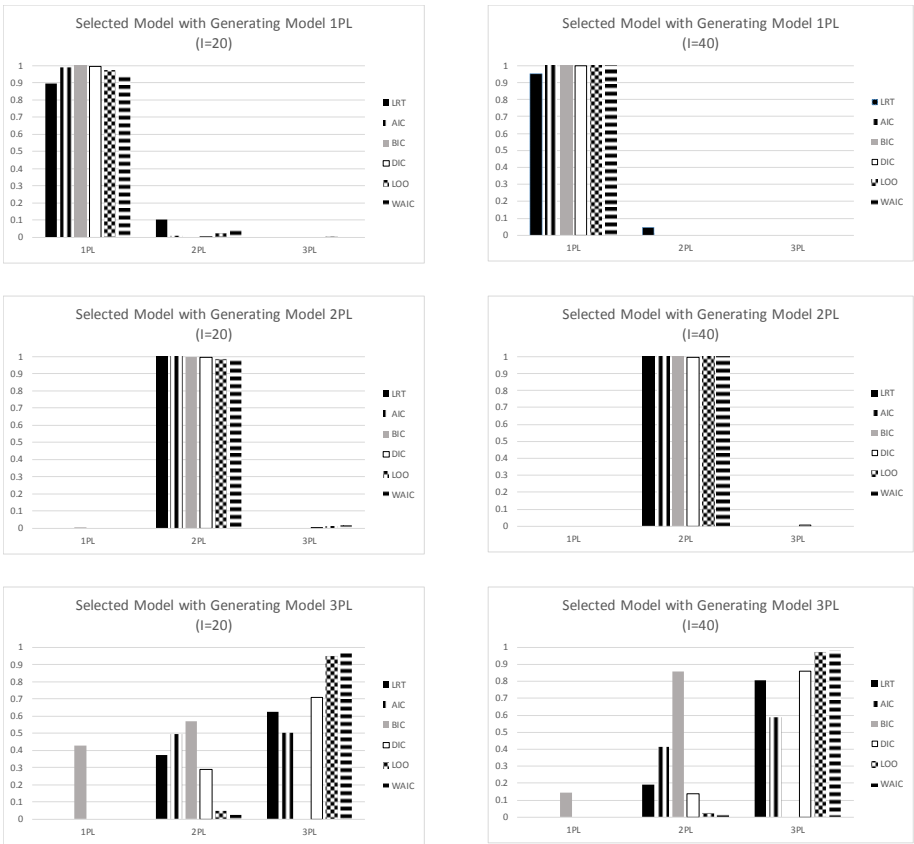

Figure 2:
Model selection by test length

although they have slightly higher power rates when the ability distribution matches the mean item difficulty. The ability distribution has a greater effect upon the performances of LRT, AIC, and DIC in that when the ability distribution does not match with the mean item difficulty, their performances drop markedly: the power rates of LRT and DIC are higher than 0.90 when the ability and difficulty match; when there is a mismatch, the power rates decrease to approximately 0.70 for DIC, 0.60 for LRT, and 0.45 for AIC.

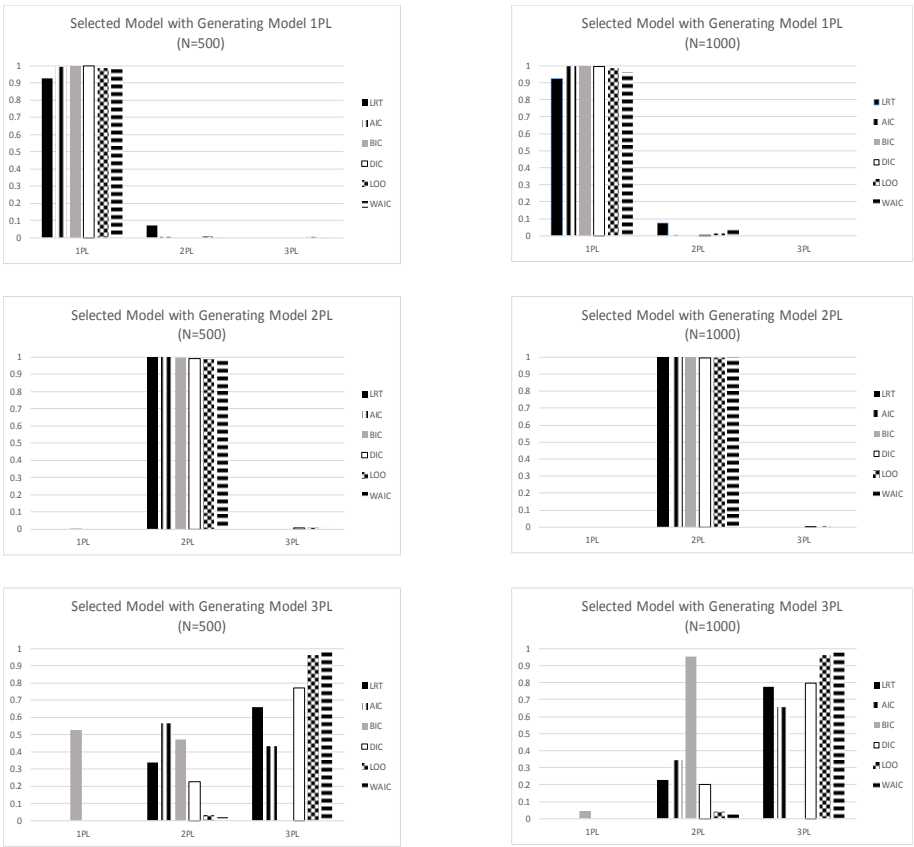

Figure 3:
Model selection by sample size

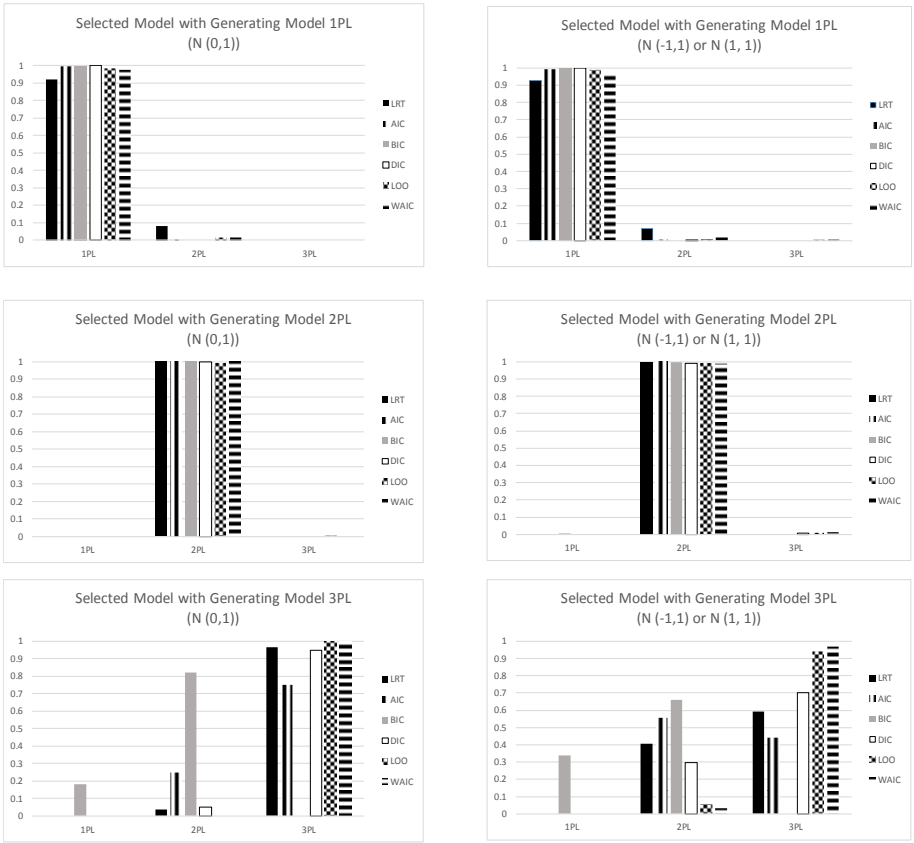

Figure 4:
Model selection by ability distribution

A real data example

In this section, we illustrate the relative performance of each of these six model selection methods with real data from a high-stakes test. In addition, we create sub-datasets by extracting different numbers of examinees and different numbers of items from the same dataset to show the potential impact of varying sample size and test length upon the six methods regarding the choice of a best fitting model.

We use a test form of the General Aptitude Test (GAT) Verbal Section (GAT-V), a high-stakes tests used for college admission purposes in Saudi Arabia. It has been shown that GAT-V is approximately unidimensional (Dimitrov & Shamrani, 2015), and we proceed with item calibration using the three unidimensional dichotomous IRT models without conducting any dimensionality assessment procedure in this example. There are 52 items in the test, and 36,886 students took the chosen test form. We randomly sample 500 (N

=500), 1,000 (N =1000), and 2,000 students (N = 2000) from the data and for each sample size, we either use the whole test with 52 items (I = 52) or the first 30 items excluding the reading comprehension items (I =30). With the resulting six data sets (three sample sizes × two test lengths) we estimate the 1PLM, 2PLM, and 3PLM using the same software programs as in the preceding simulation section.

Table 4 lists the model selection results based on the six methods with the six data sets. Note that in the LRT column, x indicates that the corresponding model is selected by LRT; for other five methods, the smallest value is highlighted in bold. As can be seen, LRT, DIC, WAIC, and LOO consistently select the 3PLM as the best fitting model regardless of the dataset. AIC selects the 2PLM model in dataset 2, where the sample size is small (N = 500) and the test length is relatively short (I = 30). In the other five datasets AIC selects the 3PLM model. Same as in the simulation study where BIC never selects the 3PLM model, here it always selects the 2PLM model as the best fitting one.

Table 4:
Model Selection Results for GAT-V with Different Sample Sizes and Test Lengths

Data	Model	Model Selection Method					
		LRT	AIC	BIC	DIC	WAIC	LOO
Dataset 1	1PLM		29873.88	30097.25	29230	29244.22	29244.15
	2PLM		29557.45	29995.77	28920	28871.08	28874.53
	3PLM	x	29529.58	30187.06	28830	28792.89	28798.19
Dataset 2	1PLM		16548.03	16678.68	16070	16074.53	16075.15
	2PLM		16354.13	16607.01	15870	15841.28	15847.38
	3PLM	x	16358.32	16737.63	15840	15820.38	15828.76
Dataset 3	1PLM		60100.95	60361.07	58870	58897.22	58897.10
	2PLM		59487.26	59997.67	58220	58170.82	58174.64
	3PLM	x	59394.25	60159.86	58050	58032.48	58035.02
Dataset 4	1PLM		33262.98	33415.12	32360	32364.33	32365.35
	2PLM		32817.46	33111.93	31840	31817.87	31826.35
	3PLM	x	32791.06	33232.76	31800	31773.59	31781.29
Dataset 5	1PLM		121827.50	122124.34	119400	119384.58	119384.31
	2PLM		120569.20	121151.70	117900	117882.11	117881.99
	3PLM	x	120333.10	121206.85	117600	117576.64	117579.05
Dataset 6	1PLM		67792.89	67966.52	65950	65976.28	65978.30
	2PLM		66952.56	67288.62	64930	64910.44	64918.46
	3PLM	x	66850.00	67354.08	64820	64789.52	64804.28

Note. Dataset 1 (N = 500, I = 52); Dataset 2 (N = 500, I = 30); Dataset 3 (N = 1000, I = 52); Dataset 4 (N = 1000, I = 30); Dataset 5 (N = 2000, I= 52); Dataset 6 (N = 2000, I = 30).

Conclusions and discussions

This study compared the performances of LRT, AIC, BIC, DIC, LOO, and WAIC in an attempt to investigate whether LOO and WAIC, which are fully Bayesian and hence theoretically superior model selection methods, perform better than the other common methods in the context of dichotomous IRT model selection. Another purpose of the current study was to investigate whether WAIC, an asymptotic approximation of LOO, performs similarly as LOO computed through PSLS approximation. The findings suggest that on average WAIC and LOO have the highest power rate (0.98) among the six methods in selecting the true dichotomous IRT model, DIC performs slightly worse with a power rate of 0.93, and BIC performs the worst with the lowest power rate of 0.67. As mentioned earlier, LRT, AIC, BIC, and DIC use point estimates in their computation while LOO and WAIC are computed based on the whole posterior distribution. Intuitively, it is expected that methods using more information (the posterior distribution) should perform better than those using less information (point estimate), and our findings corroborate such expectations: WAIC and LOO perform the best due to their full use of the posterior distribution, DIC comes the second due to its partial use of the posterior distribution, and the other methods (LRT, AIC, and BIC) that do not use the posterior distribution have the lowest statistical power in dichotomous IRT model selection. In other words, the theoretical advantage of WAIC and LOO being fully Bayesian indeed translates into superior performances regarding dichotomous IRT model selection in the current study. In terms of comparative performance between WAIC and LOO computed through PSLS, it seems that WAIC does not perform any worse than LOO since they have the same mean power rates. Although their performances differ slightly in some conditions, we observe that such differences are negligible and it is reasonable to conclude they have nearly identical performances in the current study.

Although WAIC and LOO seem to perform better than the other four methods, there are conditions in which they perform slightly worse. For example, when the generating model is 1PLM, both WAIC and LOO have a slightly higher probability of choosing a more parameterized model and hence slightly lower power. When 2PLM is the generating model, WAIC and LOO perform slightly worse, although the difference is smaller than in conditions where 1PLM is the generating model. WAIC and LOO outperform the other methods when 3PLM is the generating model. It is worth noting that with 3PLM as the generating model, BIC always chooses a less parameterized model and never selects the 3PLM model correctly.

Similar to what Kang and Cohen (2007) observed, in general when ability parameters are simulated from $N(1, 1)$ and the generating model is 3PLM, the performances of LRT, AIC, BIC, and DIC drop precipitously. Specifically, in our study the average power rate of LRT in such simulation conditions is 0.25, AIC 0.01, BIC 0, and DIC 0.41. Kang and Cohen attributed such drops in performance to that when ability parameters are simulated from $N(1, 1)$, items generated with a mean difficulty of zero become easy and the pseudo-guessing parameters in 3PLM are not accurately estimated. However, WAIC and LOO are less affected by such inaccuracies: in the same simulation conditions the average power rates of WAIC and LOO are 0.94 and 0.89 respectively.

In the real data analysis section, we showed that with different subsets created from the same data set with varying sample sizes and test lengths, AIC produces inconsistent results: when the sample size is small (N=500) and test is short (I=30), AIC chooses 2PLM; it chooses 3PLM with the other five datasets. BIC consistently chooses 2PLM, which is expected since in the simulation study section, it never chooses 3PLM even when it is the generating model. Consequently, we do not recommend AIC and BIC as model selection methods in practice when practitioners are interested in finding out which dichotomous IRT model fits their data best, especially when the sample size is small and the test length is short.

As for the comparison between WAIC and LOO, despite their identical mean power rates, there are some simulation conditions in which LOO performs slightly better and vice versa. We believe that such fluctuations are due to the limited number of replications within each condition and with a sufficiently larger number of replications, such discrepancies will eventually disappear. We conclude that WAIC and LOO perform similarly regarding dichotomous IRT model selection in the current simulated conditions. In addition, the time difference between the computation of WAIC and LOO in the R package **loo** is negligible. For researchers interested in selecting a dichotomous IRT model for their data, either method can be used.

One limitation of the current study is its limited scope. We focus on the scenario of model selection and comparison among several competing IRT models that are unidimensional and dichotomous, and it remains unknown how WAIC and LOO perform with more complex IRT models such as polytomous, multidimensional, multilevel, and mixture IRT models. Future studies should investigate the performances of WAIC and LOO as model selection methods with those more complex IRT models.

To sum up, WAIC and LOO perform well in the context of dichotomous IRT model selection, and we recommend that IRT researchers and practitioners consider them as desirable alternatives to the more traditional methods such as AIC and DIC, especially when items have low difficulty relative to examinees' ability and hence the pseudo-guessing parameter cannot be accurately estimated. In addition, as demonstrated by Luo and Jiao (2017), the powerful and efficient HMC algorithm implemented in Stan, which can be accessed through the R package **rstan**, allows quick estimation of dichotomous IRT models; the estimation results can be directly accessed by the R package **loo**, which provides an easy and free solution for the computation of WAIC and LOO.

Reference

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory*, 267-281. Budapest, Hungary: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*(3), 251-269.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*(4), 541-561.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*(4), 381-409.

Bolt, D. M. and Lall, V. 2003. Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*, 395–414.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345-370.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73*(2), 209-230.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software.*

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. doi:10.18637/jss.v048.i06

Cohen, A. S., & Cho, S. J. (2016). Information Criteria. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume Two: Statistical Tools (Vol.21)* (pp. 363-378). Boca Raton, FL: CRC Press.

Dimitrov, D. M., & Shamrani, A. R. (2015). Psychometric Features of the General Aptitude Test – Verbal Part (GAT-V) A Large-Scale Assessment of High School Graduates in Saudi Arabia. *Measurement and Evaluation in Counseling and Development*, *48*(2), 79-94.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153-160.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 147-167. Oxford University Press.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733-760.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457-472.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997-1016.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement, 20*(2), 101-125.

Hickendorff, M., Heiser, W. J., Van Putten, C. M., & Verhelst, N. D. (2009). Solution strategies and achievement in Dutch complex arithmetic: Latent variable modeling of change. *Psychometrika*, *74*(2), 331-350.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593-1623.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297-307.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, *17*(2), 295-311.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331-358.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*(7), 499-518.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Jiao, H., Wang, S., & He, W. (2013). Estimation Methods for One-Parameter Testlet Models. *Journal of Educational Measurement, 50*(2), 186-203.

Jiao, H., & Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology, 68*(1), 65-83.

Levy, R. and Mislevy, R.J. (2016). *Bayesian Psychometric Modeling*. Boca Raton, FL: Chapman and Hall/CRC.

Levy, R., Mislevy, R. J., and Sinharay, S. 2009. Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*(7), 519-537.

Li, T., Xie, C., & Jiao, H. (2016, May 30). Assessing Fit of Alternative Unidimensional Polytomous IRT Models Using Posterior Predictive Model Checking. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000082

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3-21.

Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*(5), 353-373.

Luo, Y., & Jiao, H. (2017). Using the Stan Program for Bayesian Item Response Theory. *Educational and Psychological Measurement*, 0013164417693666.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

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*(1), 63-79.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159-76.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*, 113-162.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 99-138. Retrieved from http://www.jstor.org/stable/2346088

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral statistics*, *24*(4), 342-366.

Revuelta, J. (2008). The generalized logit-linear item response model for binary-designed items. *Psychometrika*, *73*(3), 385-405.

Rijmen, F., De Boeck, P., & Leuven, K. U. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*(3), 271-285.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation*, *72*(3), 217-232.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika, 37*(2), 87-110.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*(4), 375-394.

Sinharay S. (2016). Bayesian model fit and model comparison. In van der Linden W. (Ed.), *Handbook of item response theory: Vol. 2. Statistical tools*. Boca Raton, FL: Chapman & Hall/CRC Press.

Sinharay, S., Johnson, M. S., and Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298-321.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Shibata, R. (1989). Statistical aspects of model selection. In J. C. Willems (Ed.), *From data to modeling* (pp. 216-240). Berlin: Springer-Verlag.

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 (Statistical Methodology)*, *64*(4), 583-639.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2010). *OpenBUGS Version 3.1.1 User Manual*. Helsinki, Finland. Retrieved from http://www.openbugs.info/.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B: Statistical Methodological, 39,* 44-47.

Vehtari, A., Gelman, A., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646.*

Vehtari, A., Gelman, A., & Gabry, J. (2016a). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing,* 1-20.

Vehtari, A., Gelman, A., and Gabry, J. (2016b). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 0.1.6.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(Dec), 3571-3594.

Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The performance of IRT model selection methods with mixed-format tests. *Applied Psychological Measurement*, *36*(3), 159-180.

Whittaker, T. A., Chang, W., & Dodd, B. G. (2013). The Impact of Varied Discrimination Parameters on Mixed-Format Item Response Theory Model Selection. *Educational and Psychological Measurement*, *73*(3), 471-490.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*(6), 469-492.

Zhu, X., & Stone, C. A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *Journal of Educational Measurement, 48*(1), 81-97.

Zhu, X., & Stone, C. A. (2012). Bayesian comparison of alternative graded response models for performance assessment applications. *Educational and Psychological Measurement*, *72*(5), 774-799.