Behaviormetrika (2023) 50:93–120
https://doi.org/10.1007/s41237-022-00167-x

ORIGINAL PAPER

Empirical evaluation of fully Bayesian information criteria for mixture IRT models using NUTS

Rehab AlHakmani[1] · **Yanyan Sheng**[2]

Received: 15 April 2021 / Accepted: 4 May 2022 / Published online: 17 June 2022

Abstract
This study is to evaluate the performance of fully Bayesian information criteria, namely, LOO, WAIC and WBIC in terms of the accuracy in determining the number of latent classes of a mixture IRT model while comparing it to the conventional model via non-random walk MCMC algorithms and to further compare their performance with conventional information criteria including AIC, BIC, CAIC, SABIC, and DIC. Monte Carlo simulations were carried out to evaluate these criteria under different situations. The results indicate that AIC, BIC, and their related CAIC and SABIC tend to select the simpler model and are not recommended when the actual data involve multiple latent classes. For the three fully Bayesian measures, WBIC can be used for detecting the number of latent classes for tests with at least 30 items, while WAIC and LOO are suggested to be used together with their effective number of parameters in choosing the correct number of latent classes.

Keywords Mixture IRT models · Information criteria · No-U-Turn sampler · LOO · WAIC · WBIC

1 Introduction

Conventional unidimensional item response theory (IRT) models assume that the observed response data stem from a homogenous population of individuals. However, under many test situations, and especially in situations where a mixture of several latent classes (i.e., subpopulations) is involved, fitting a conventional IRT model

Communicated by Maomi Ueno.

✉ Rehab AlHakmani
rehab.alhakmani@ecae.ac.ae

Yanyan Sheng
y.sheng@uchicago.edu

[1] Emirates College for Advanced Education, Abu Dhabi, UAE

[2] The University of Chicago, Chicago, IL 60637, USA

to the data produces biased estimates of model parameters (e.g., De Ayala et al. 2002). As a result, mixture IRT (MixIRT) models (Rost 1990) were developed to capture the presence of these latent classes that are qualitatively different but within which a conventional IRT model holds. De Ayala et al. (2002) further showed how the occurrence of differential item functioning (DIF) items can be explained by realizing that the data do not come from a homogenous population of individuals but are a mixture of multiple populations. In the MixIRT modeling framework, persons are characterized by their location on a continuous latent dimension as well as by their latent class membership. Also, each subpopulation has a unique set of item parameters.

Different estimation methods have been developed to estimate IRT models, with the current focus on the fully Bayesian estimation based on Markov chain Monte Carlo (MCMC; Hastings 1970; Metropolis and Ulam 1949; Metropolis et al. 1953) techniques. Researchers have documented its advantages over the traditional maximum likelihood estimation (MLE; Fisher 1922) methods in estimating various IRT models (e.g., Finch and French 2012; de la Torre et al. 2006; Wollack et al. 2002). The MLE method may result in infinite or implausible parameter estimates in situations where unusual response patterns are encountered such as perfect or zero scores. On the other hand, the fully Bayesian estimation via the use of the MCMC simulation techniques approximates the joint posterior distribution of all model parameters, and hence accounts for the uncertainty associated with any parameter estimation. However, using MCMC can be time consuming and computationally expensive. Recent developments of MCMC focus on non-random walk MCMCs such as the no-U-turn sampler (NUTS; Hoffman and Gelman 2014), which avoids the inefficient exploration of the parameter space via random walks. Consequently, NUTS converges to the posterior distribution faster than the common MCMC algorithms such as Gibbs sampling (Geman and Geman 1984) and Metropolis–Hastings (MH; Hastings 1970; Metropolis and Ulam 1949) even for complex models such as MixIRT models. Indeed, Luo and Jiao (2017) showed how NUTS was efficient in fitting various IRT models, namely the three-parameter logistic (3PL; Birnbaum 1968) IRT model, the graded response model (GRM; Samejima 1969), and the nominal response model (NRM; Bock 1972). For example, they found that around 200 iterations were sufficient to reach convergence for dichotomous IRT models (including the one-, two-, and three-parameter logistic models). Uto and Ueno (2020) also demonstrated the advantage of NUTS in estimating a complex generalized many-facet Rasch model that simultaneously incorporates three-rater characteristic parameters, including severity, consistency, and range restriction, especially for data with relatively small sample sizes.

Previous work on estimating mixture IRT models using MCMC algorithms focused mainly on implementing Gibbs sampling (e.g., Cho et al. 2013). In an effort of applying NUTS to a two-parameter MixIRT model, Al Hakmani and Sheng (2019) demonstrated the accuracy of NUTS in recovering model parameters and class membership of individual persons although the recovery of the class membership was not satisfactory for test situations where more than two classes were involved. Given the efficiency of NUTS, this study focuses on using it to evaluate three fully Bayesian information criteria, namely, the widely applicable (or

Watanabe–Akaike) information criterion (WAIC; Watanabe 2010), the widely applicable Bayesian information criterion (WBIC; Watanabe 2013), and the leave-one-out cross-validation (LOO) implemented through a Pareto smoothed important sampling (LOO-PSIS; Vehtari et al. 2017), in the context of MixIRT models in terms of the accuracy in determining the number of latent classes, and further to compare the performance of these fully Bayesian information criteria with other information criteria.

In the literature, various forms of information criteria, under either the frequentist or the Bayesian framework, have been used to assess the fit of conventional IRT and MixIRT models, including the popular Akaike's information criterion (AIC; Akaike 1974), Bayesian information criterion (BIC; Schwarz 1978), and deviance information criterion (DIC; Spiegelhalter et al. 2002). Two adjusted forms of AIC and BIC, namely, consistent AIC (CAIC; Bozdogan 1987) and sample-size-adjusted BIC (SABIC; Sclove 1987), have been less commonly used in detecting the number of latent classes in the MixIRT literature.

While Vehtari et al. (2017) noted that the above-referenced criteria are not fully Bayesian and instead recommended the use of the WAIC and the LOO-PSIS indices, previous research has mainly focused on using them to evaluate their performance in fitting MixIRT models in the fully Bayesian framework. Specifically, Choi et al. (2017) investigated the performance of AIC, BIC, corrected AIC (AICc; Sugiura 1978), and SABIC in detecting the correct number of latent classes in the two-class mixture Rasch model. Their results revealed that the four information criteria performed differently under different class-distinction conditions with the overall conclusion that AICc and SABIC performed comparable to or better than AIC and BIC. Also, Lee and Beretvas (2014) evaluated the performance of AIC, BIC, CAIC, and SABIC in identifying the correct mixture Rasch model while manipulating DIF effect sizes and latent class proportions. Their findings indicated that these information criteria performed better with equal class proportions than unequal class proportions, and that for the condition of equal class proportions and small DIF effect sizes, the AIC performed better than other criteria in selecting the correct model. Alternatively, Li et al. (2009) examined the performances of AIC, BIC, and DIC in selecting the correct MixIRT model among three competing models (the mixture one-, two- and three-parameter logistic IRT models) via the use of Gibbs sampling, and found that BIC was the most effective, while AIC tended to choose more complex models in certain conditions and DIC was the least effective method. Similarly, Nylund et al. (2007) examined the performance of four information criteria including AIC, BIC, CAIC, and SABIC in detecting the number of classes for three mixture models (the latent class analysis, the factor mixture model, and the mixture growth model), and concluded that BIC and SABIC were better than AIC in identifying the correct number of classes and that this performance improved as the total sample size increased. They further noted that CAIC performed well in correctly identifying the number of classes for most conditions except for the most complicated model (i.e., the 10-item categorical latent class analysis model with a complex structure and unequal class sizes). Sen et al. (2019), while examining the performance of AIC, BIC, CAIC, and SABIC in detecting the best fitting

multivariate mixture Rasch model with a varying number of classes at the student level and school level, concluded that CAIC and BIC performed better than AIC or SABIC in identifying the correct model.

The performance of fully Bayesian measures such as WAIC and LOO has been investigated in the context of unidimensional IRT models. For example, Luo and Al-Harbi (2017) compared their performances with four popular methods: the likelihood ratio test (LRT; Neyman and Pearson 1933), AIC, BIC, and DIC, for fitting dichotomous IRT models including the conventional one-, two-, and three-parameter logistic IRT models. Their results showed that WAIC and LOO performed consistently better than the other four criteria, especially for fitting the 3PL model. For polytomous IRT models [e.g., the graded response model (GRM; Samejima 1969), the rating scale model (RSM; Andrich 1978), the partial credit model (PCM; Masters 1982), and the generalized partial credit model (GPCM; Muraki 1992)], Luo (2019) compared the performances of WAIC and LOO with AIC, BIC, AICc, SABIC, and DIC and found that all the seven measures had relatively high statistical power in detecting the true polytomous IRT model with the frequentist-based measures (AIC, BIC, AICc, and SABIC) performing slightly better than the Bayesian ones (DIC, LOO, and WAIC). Moreover, da Silva et al. (2018) suggested to employ DIC, WAIC, and LOO over expected AIC (Brooks et al. 2002) or expected BIC (Carlin and Louis 2001) in fitting polytomous IRT models (GRM and GPCM), especially for data with small sample sizes (e.g., 50–150 persons) and short tests (e.g., 7–15 items). Nevertheless, to date, WAIC, LOO or WBIC has not been considered or evaluated for MixIRT models. It is hence necessary to explore their usage under the MixIRT framework in identifying the number of latent classes.

In view of the above, this study is to assess the performance of fully Bayesian information criteria, namely LOO, WAIC, and WBIC in terms of the accuracy in determining the number of latent classes of a MixIRT model by comparing it to the conventional IRT model. For the sake of comparisons, AIC, BIC, CAIC, SABIC, and DIC were considered in this study. The rest of the paper is organized as follows. Section 2 briefly describes a form of the MixIRT model with its prior specifications as implemented in NUTS, followed by Sect. 3 where we describe all the information criteria considered in this study. In Sect. 4, a Monte Carlo simulation study is presented to evaluate the performance of the three fully Bayesian measures in fitting a MixIRT model via NUTS while comparing them with other measures. Its results are summarized in Sect. 5 for tests with multiple latent populations and for those with a single population, with a few remarks discussed in Sect. 6.

2 Model and prior specifications

This study focuses on evaluating the three fully Bayesian information criteria in detecting the number of latent classes of the mixture two-parameter logistic (Mix2PL) model via implementing NUTS and further on comparing them with other information criteria. The model and its prior specifications as implemented in NUTS are briefly described in this section.

2.1 Two-parameter mixture IRT model

In the Mix2PL model, the conventional 2PL IRT model is assumed to hold for each latent class, allowing the item difficulty and discrimination parameters to differ for different classes. Moreover, each person is parameterized by a class membership parameter g and a class-specific ability parameter θ_{ig}, whereas each item is parameterized by a different set of difficulty and discrimination parameters for each latent class. The probability of a correct ($Y_{ij}=1$) response for person i to item j in the Mix2PL IRT model is defined as

$$P\left(Y_{ij}=1\right)=\sum_{g=1}^{G}\pi_g\times\frac{\exp\left[a_{jg}\left(\theta_{ig}-b_{jg}\right)\right]}{1+\exp\left[a_{jg}\left(\theta_{ig}-b_{jg}\right)\right]},i=1,\ldots,n,j=1,\ldots,J, \quad (1)$$

where $g=1, 2,\ldots, G$ is the latent class indicator; b_{jg} and a_{jg} denote the difficulty and discrimination parameters, respectively, for item j in the gth class; θ_{ig} denotes the ability for person i who belongs to class g; and π_g denotes the proportion of persons in each class (i.e., the mixing proportion) such that these proportions sum to one. In situations where there is only one latent class (i.e., $G=1$), the Mix2PL model shown in Eq. (1) is reduced to the conventional 2PL model as defined in Eq. (2):

$$P\left(Y_{ij}=1\right)=\frac{\exp[(a_j\left(\theta_i-b_j\right)]}{1+\exp[(a_j\left(\theta_i-b_j\right)},i=1,\ldots,n,j=1,\ldots,J, \quad (2)$$

where θ_i is the ability of person i, and b_j and a_j are difficulty and discrimination parameters, respectively, for item j.

2.2 NUTS and prior specification

In the fully Bayesian framework, common MCMC algorithms such as Gibbs sampling (Geman and Geman 1984) and Metropolis–Hastings (MH; Hastings 1970; Metropolis and Ulam 1949) explore the posterior distribution via simple random walk proposals, and as a result, a large number of iterations are needed to sufficiently explore the parameter space. Conversely, non-random walk MCMC algorithms such as Hamiltonian Monte Carlo (HMC; Duane et al. 1987; Neal 2011) and the no-U-turn sampler (NUTS; Hoffman and Gelman 2014) avoid the inefficient exploration of the parameter space. HMC is a powerful tool, but its performance depends on choosing suitable values for the step size parameter ε and the number of leapfrog steps L (Hoffman and Gelman 2014). Tuning these parameters, and specifically L requires some expertise and preliminary runs (Hoffman and Gelman 2014; Neal 2011). NUTS is an extension of HMC that eliminates the need to specify the number of leapfrog steps parameter L. This is achieved using a criterion based on the dot product between the current momentum $\widetilde{r}$ and the distance between the proposal and the initial value of the model parameter ξ, which is proportional to the progress one would make away from the starting point ξ. This algorithm in which one runs

leapfrog steps until $(\widetilde{\xi} - \xi) \bullet \widetilde{r}$ is less than 0, however, does not ensure time reversibility and hence convergence to the correct distribution is not ensured. This issue is resolved using a recursive algorithm in which NUTS creates a set of candidate values that spans a wide path of the target distribution, stopping automatically when it starts to make a U-turn (Hoffman and Gelman 2014), at which point NUTS stops the simulation and samples from the set of values computed during the simulation. In practice, NUTS performs as efficient as, and sometimes better than, a well-tuned HMC without requiring user interventions.

To implement NUTS for the Mix2PL model in this study, priors and hyperpriors have been specified to be similar to those used by others (e.g., Meyer 2010; Li et al. 2009) such that normal prior densities were assumed for person ability parameters $\theta_{ig} \sim N(\mu_g, 1)$, with a standard normal distribution for the hyperparameters μ_g, and a Dirichlet distribution for the mixing-proportion parameters $(\pi_1, \ldots, \pi_G) \sim \text{Dirichlet}(1, \ldots, 1)$; a standard normal prior was considered for the class-specific difficulty parameters, and a truncated normal prior for the class-specific discrimination parameters $a_{jg} \sim N_{(0,\infty)}(0, 1)$. In addition, the sum-to-one constraint on the mixing proportions was achieved by assigning the mixing proportions a unit simplex, and the problem of label switching was resolved by imposing an ordinal constraint on the mean ability (μ_g) parameters and the item difficulty parameters (b_g).

3 Information criteria

Information criteria are statistical measures for the comparative evaluation among candidate models, and they provide ways to assess a model based on its log-likelihood and complexity. Among the various measures considered in the IRT literature, AIC, BIC, CAIC, and SABIC are frequentist information criteria that have been developed based on the MLE method, and can be calculated as

$$\text{AIC} = -2\log(L) + 2d, \tag{3}$$

$$\text{BIC} = -2\log(L) + d\big[\log(n)\big], \tag{4}$$

$$\text{CAIC} = -2\log(L) + d\big[\log(n) + 1\big], \tag{5}$$

$$\text{SABIC} = -2\log(L) + d\left[\log\left(\frac{n+2}{24}\right)\right], \tag{6}$$

where d is the number of estimated parameters, $\log(L)$ is the natural logarithm of the likelihood function obtained from the MLE, and n is the sample size. These information criteria may provide different solutions to the same data due to differences in the penalty function applied to the likelihood. It is noted that in the literature, AIC is argued to be an inconsistent measure, and that given consistency is an asymptotic property expected from a model-selection method, Bozdogan (1987) proposed CAIC, an asymptotically consistent measure that includes a penalty for

models with larger numbers of parameters using the sample size n. In addition, BIC applies a penalty term that uses both the number of parameters and the sample size, and has been found to be somewhat more accurate than AIC for selection of MixIRT models (Li et al. 2009; Preinerstorfer and Formann 2012). Sugiura (1978) introduced a sample-size-adjusted form of BIC (SABIC) that replaces n in the BIC equation with $(n+2)/24$. Similar to BIC, the penalty for adding more parameters is represented by both the number of parameters and the sample size. The penalty term, however, is not as large as that in BIC. These frequentist measures are suitable when model parameters are estimated using the MLE estimation method. Nevertheless, they can be applied to the fully Bayesian setting via an approach described by Congdon (2003), where the MLE-based deviance value, $-2\log(L)$, is replaced with the posterior mean of the deviance $\overline{D(\xi)}$ obtained using an MCMC algorithm (Congdon 2003) such as NUTS, in which ξ denotes all model parameters. Congdon (2003) pointed out the limitation of the frequentist measures in likelihood comparisons of discrete mixture models involving varying numbers of classes while the process involved in Bayesian model selection is simpler and has advantages in comparing non-nested models. This approach has been adopted by multiple studies that implemented a random-walk MCMC algorithm (i.e., Gibbs sampler) in the mixture IRT literature either to evaluate the performance of, e.g., AIC and BIC in detecting the correct MixIRT model (e.g., Li et al. 2009; Sen et al. 2019) or to inform model selection (Cho et al. 2013; Sen et al. 2016), and consequently, is considered by this study.

In the context of Bayesian estimation, Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC), a Bayesian counterpart of the AIC measure that is defined as

$$\text{DIC} = \overline{D(\xi)} + p_{\text{DIC}}, \tag{7}$$

where p_{DIC} is the effective number of parameters defined as

$$p_{\text{DIC}} = \overline{D(\xi)} - D(\widehat{\xi}), \tag{8}$$

where $D(\widehat{\xi})$ is the deviance obtained from the posterior estimates of the parameters. Given that $D(\widehat{\xi})$ is based on a point estimate instead of the entire posterior distribution, DIC is not considered as a fully Bayesian measure. Vehtari et al. (2017) specifically pointed out that its effective number of parameters for a model can result in negative values, while Plummer (2008) noted to the argument in the literature between the advantages of DIC in practice and the lack of a clear theoretical foundation.

AIC and BIC can be used for assessing statistical models when such models are regular, and the likelihood function can be approximated by a normal distribution (Watanabe 2021). For statistical models with a hierarchical structure or with latent variables (i.e., singular models), where regularity is not met, two fully Bayesian information criteria, namely WAIC (Watanabe 2010) and WBIC (Watanabe 2013), were proposed as generalizations of AIC and BIC, respectively. These information criteria estimate the generalization loss (i.e., the average minus log predictive likelihood) and the Bayes free

energy (i.e., the minus logarithm of Bayes marginal likelihood), respectively (Watanabe 2021). WBIC is specifically defined as

$$\mathrm{WBIC} = E_{\xi}^{\beta}\left[nL_n(\xi)\right], \tag{9}$$

where $\beta = 1/\log(n)$ is the inverse temperature and $E_{\xi}^{\beta}[G(\xi)]$, for an arbitrary function $G(\xi)$, is the expectation over the posterior distribution under the inverse temperature β, defined as follows

$$E_{\xi}^{\beta}[G(\xi)] = \frac{\int G(\xi)\prod_{i=1}^{n} p(y_i|\xi)^{\beta}\varphi(\xi)\mathrm{d}\xi}{\int \prod_{i=1}^{n} p(y_i|\xi)^{\beta}\varphi(\xi)\mathrm{d}\xi}, \tag{10}$$

where $\varphi(\xi)$ denotes the prior probability density function of a given model parameter $\xi \subset \Xi$. It is noted that for regular models, WBIC reduces to BIC (Watanabe 2013). Watanabe (2021) further described the mathematical foundation of WBIC and discussed its application to a normal mixture model, a common singular model.

In addition, WAIC and LOO are two fully Bayesian information criteria that estimate the predictive accuracy of the fitted model using available data, without waiting for out-of-sample data (Gelman et al. 2014). WAIC estimates the out-of-sample expectation by first computing a log pointwise posterior predictive density (LPPD) of the data, which is defined as

$$\mathrm{LPPD} = \sum_{i=1}^{n} \log \int p(y_i|\xi)p_{\mathrm{post}}(\xi)\mathrm{d}\xi, \tag{11}$$

where $p_{\mathrm{post}}(\xi) = p(\xi|y)$ is the posterior distribution of model parameters ξ. In practice, LPPD is computed by evaluating the expectation via sampling from the posterior distribution $p_{\mathrm{post}}(\xi)$ such that

$$\mathrm{LPPD} = \sum_{i=1}^{n} \log\left[\frac{1}{S}\sum_{s=1}^{S} p(y_i|\xi^s)\right], \tag{12}$$

where $s = 1, 2, \ldots, S$ denotes the number of simulation samples from the posterior density. After computing the LPPD, WAIC is obtained by adding a correction (p_{WAIC}) for the effective number of parameters to adjust for overfitting

$$\mathrm{WAIC} = -2\mathrm{LPPD} + 2p_{\mathrm{WAIC}}. \tag{13}$$

The correction term, p_{WAIC}, can be computed in the following two ways:

$$p_{\mathrm{WAIC1}} = 2\sum_{i=1}^{n} \log\left(E_{\mathrm{post}}p(y_i|\xi)\right) - E_{\mathrm{post}}\left(\log\left(p(y_i|\xi)\right)\right), \tag{14}$$

$$p_{\mathrm{WAIC2}} = \sum_{i=1}^{n} \mathrm{var}_{\mathrm{post}}\left[\log\left(p(y_i|\xi)\right)\right]. \tag{15}$$

The second adjustment as expressed in Eq. (15) is more computationally stable since summing the variance for each data point produces stability (Gelman et al. 2014), and is implemented in the R package *loo* (Vehtari et al. 2017), which is used for computation of both WAIC and LOO. LOO, on the other hand, is tied with Bayesian cross-validation studies, where a dataset is repeatedly partitioned into a training set and a validation set. The model of interest is fitted to the training set to obtain the posterior distribution, with which the fit of the model to the validation set is evaluated. LOO is a special case of cross-validation in which one data point is left out each time and the LPPD is computed with the remaining data points as follows

$$\text{LPPD}_{\text{LOO}} = \sum_{i=1}^{n} \log \int p(y_i|\xi) p_{\text{post}(-i)}(\xi) \text{d}\xi, \tag{16}$$

where $p_{\text{post}(-i)}(\xi)$ is the posterior distribution without the ith data point, and is computed as

$$\text{LPPD}_{\text{LOO}} = \sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(y_i|\xi^{is}) \right], \tag{17}$$

where ξ^{is} is the sth simulated value from the posterior distribution conditioning on the dataset without the ith data point (Gelman et al. 2014). To place LOO on the same scale as WAIC, the computed LPPD_{LOO} is multiplied by -2. According to Gelman et al. (2014), WAIC is asymptotically equal to LOO. To obtain more accurate estimate of LOO, Vehtari et al. (2017) suggested fitting a Pareto distribution to the upper tail of the distribution of the importance weights (PSIS) and argued that PSIS smoothing approach would benefit from using stable importance weights. LOO, WAIC, and WBIC are less used in practice due to the requirement of additional computational steps. In spite of the computational expense, they have advantages over simpler estimates in terms of being fully Bayesian measures that use the whole posterior distribution in contrast to point estimates such as AIC or DIC (Vehtari et al. 2017). Unlike simpler estimates, these fully Bayesian information criteria can be used for singular models with hierarchical and mixture structures or for models with different prior specifications, and hence are expected to perform better than point estimate-based information criteria in selecting or comparing singular complex models (Watanabe 2013) such as MixIRT models.

4 Method

Monte Carlo simulations were carried out to investigate the performance of the three fully Bayesian information criteria in recovering the number of latent classes, via fitting the Mix2PL model using NUTS and further in comparing with other information criterion measures when one or multiple latent classes exist. To ensure all Markov chains have converged to their stationary distributions, the number of warm-up iterations that should be discarded and the number of sampling iterations that should be used to estimate the posterior distribution were determined using the

Gelman-Rubin R statistic (Gelman and Rubin 1992) with a threshold of 1.10 as suggested by Gelman et al. (2014), as non-convergence can result in inaccurate estimation of model parameters for MixIRT models (Jang and Cohen 2020), which in turn can affect the model-data fit.

In the MixIRT literature, the sample size, the test length, and the number of latent classes appear to affect parameter recovery of the MixIRT model. For instance, Preinerstorfer and Formann (2012) found that increasing both the sample size (500, 1000, 2500) and the number of items (10, 15, 25, 40) led to higher accuracy in estimating parameters of the mixture Rasch model. Moreover, Li et al. (2009) found that recovery of item difficultly and discrimination parameters in different MixIRT models (i.e., mixture one-, two-, or three-parameter logistic models) differed based on the number of latent classes (1, 2, 3, 4), test lengths (6, 15, 30), and sample sizes (600, 1200). Difficulty and discrimination parameters were mostly affected by the number of latent classes such that when the number of latent classes increased, the recovery of model parameters was less accurate. Also, their results indicated that the root mean square error (RMSE) decreased as sample size and test length increased. The percentage of correct classifications of class membership for individual persons increased with an increase in test length. Different sets of sample size and test length conditions have been used in the MixIRT literature. For instance, Bilir (2009) and Samuelsen (2005) simulated sample sizes of 500 and 2000 with 20 items, while Meyer (2010) simulated the same sample sizes but with 25 items.

Equal mixing proportions were considered by many studies to simulate test data involving multiple latent classes for different purposes (e.g., Bolt et al. 2001, 2002; Meyer 2010; Cho et al. 2013). For example, Bolt et al. (2001) as well as Cho et al. (2013) set the mixing proportions for each latent class to be equal. Specifically, they set $\pi = (0.50, 0.50)$ in the two-class condition, (0.33, 0.33, 0.33) in the three-class condition, and (0.25, 0.25, 0.25, 0.25) in the four-class condition. Similarly, Meyer (2010) and Bolt et al. (2002) specified mixing proportions of $\pi = (0.50, 0.50)$ for the speeded class and the non-speeded class. Further, Preinerstorfer and Formann (2012) found that item parameters were recovered more precisely in the condition of equally sized subgroups (i.e., $\pi = 0.50, 0.50$).

In the Monte Carlo simulations, two sets of test conditions were considered, with the first set treating the two-class Mix2PL model as the true model whereas the second set treating the conventional 2PL model as true. Simulation conditions were considered based on prior studies to reflect some practical considerations. Specifically, with binary item response data generated from each set of conditions for sample sizes ($n = 500$ and 1000) and test lengths ($J = 15$, 20, 30), NUTS was implemented to fit three candidate models, namely the 2PL model, the two-class Mix2PL model, and the three-class Mix2PL model (i.e., $G = 1$, 2 or 3). For the first set of test conditions, data were generated using the two-class Mix2PL model, as defined in Eq. (1), with equal mixing proportions (i.e., $\pi_1 = 0.50$ and $\pi_2 = 0.50$). The model parameters were generated such that the person ability parameters were generated from a mixture of two subpopulations where $\theta_1 \sim N(-2, 1)$ and $\theta_2 \sim N(2, 1)$; the class-specific item difficulty parameters were generated from a uniform distribution where $b_1 \sim U(-2, 0)$ and $b_2 \sim U(0, 2)$; and the class-specific item discrimination parameters were generated from a uniform distribution where $a_g \sim U(0, 2)$, $g = 1$ or

2. For the second set of test conditions, data were generated using the 2PL model, as defined in Eq. (2). The model parameters were generated such that the person ability parameters were generated from a standard normal distribution $\theta \sim N(0, 1)$; the item difficulty parameters were generated from a uniform distribution where $b \sim U(-2, 2)$; and the item discrimination parameters were generated from a uniform distribution where $a \sim U(0, 2)$.

To assess the recovery of the number of latent classes for each data set, the three fitted models were compared using the three fully Bayesian information criteria, namely WAIC, LOO, and WBIC, as well as other measures including AIC, BIC, CAIC, SABIC, and DIC. After computing the respective information criterion measure for each of the three candidate models, the model with the smallest value was selected as the best fitting model. With 50 replications, the proportion of the time the generating model was selected as the best fitting model indicates the accuracy of recovering the number of latent classes by each information criterion. In addition, the values of the eight information criteria were further averaged across replications to provide summary information. To increase the efficiency of computations, the simulations were carried out on a high-performance computing cluster that includes a total of 16,016 cores across 572 nodes and 2.2 PB of storage.

RStan (Stan Development Team 2020) was used to implement NUTS to fit the IRT models, and the Stan code for computing the posterior distribution under the inverse temperature $\beta = 1/\log(n)$ is provided in Appendix 1. This code was used for fitting the conventional 2PL model and two- and three-class Mix2PL models before computing WBIC using R. The Stan code for computing the standard posterior distribution when the inverse temperature $\beta = 1$ is presented in Appendix 2. This code is similar to that in Appendix 1 except that the log-likelihood is multiplied by 1, and it was used for fitting the same models before computing the other information criteria, AIC, BIC, CAIC, SABIC, DIC, LOO, and WAIC in R.

5 Results

With four Markov chains, 1000–4000 warm-up iterations and 4000–12,000 sampling iterations were used to ensure the convergence of each NUTS implementation, and specifically that the Gelman–Rubin R statistic was less than the recommended threshold of 1.10 for each model parameter in each simulated condition.

5.1 Results for tests with subpopulations

The results for the first six conditions where data conformed to the two-class Mix2PL model are summarized in Tables 1, 2, 3 and 4 with Tables 1 and 2 displaying the number and proportion of the time each model was identified as the best-fitting model according to each criterion (where the numbers in bold are for the model with the highest frequency/proportion among the three candidate models), and Tables 3 and 4 show the average information criteria and effective number of parameters averaged across the 50 replications (where the smallest average

Table 1 Frequency (relative frequency) for selecting each candidate models where data conformed to the 2-class Mix2PL model ($n = 500$)

J	Candidate model	Model selection method							
		LOO	WAIC	WBIC	DIC	AIC	BIC	CAIC	SABIC
15	2PL	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	**26 (0.52)**	**34 (0.68)**	0 (0.00)	**38 (0.76)**	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	24 (0.48)	16 (0.32)	**50 (1.00)**	12 (0.24)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)
20	2PL	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	19 (0.38)	**30 (0.60)**	8 (0.16)	**37 (0.74)**	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	**31 (0.62)**	20 (0.40)	**42 (0.84)**	13 (0.26)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)
30	2PL	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	19 (0.38)	**21 (0.42)**	**49 (0.98)**	**41 (0.82)**	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	**31 (0.62)**	19 (0.58)	1 (0.02)	9 (0.18)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)

The maximum frequency of selecting a model is 50. Values in bold represent the largest frequency out of the 50 replications

Table 2 Frequency (relative frequency) for selecting each candidate models where data conformed to the 2-class Mix2PL model ($n = 1000$)

J	Candidate model	Model selection method							
		LOO	WAIC	WBIC	DIC	AIC	BIC	CAIC	SABIC
15	2PL	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	25 (0.50)	**36 (0.72)**	0 (0.00)	**42 (0.84)**	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	25 (0.50)	14 (0.28)	**50 (1.00)**	8(0.16)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)
20	2PL	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	**28 (0.56)**	**35 (0.70)**	7 (0.14)	**47 (0.94)**	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	22 (0.44)	15 (0.30)	**43 (0.86)**	3 (0.06)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)
30	2PL	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	**49 (0.98)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	21 (0.42)	**28 (0.56)**	**48 (0.96)**	**44 (0.88)**	1 (0.02)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	**29 (0.58)**	22 (0.44)	2 (0.04)	6 (0.12)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)

The maximum frequency of selecting a model is 50. Values in bold represent the largest frequency out of the 50 replications

values among the three models are marked in bold). An inspection of Tables 1 and 2 reveals the following findings:

- Although AIC occasionally (1 out of 50 or 2% of the time) favored the correct two-class Mix2PL model when $n = 1000$ and $J = 30$, the four frequentist infor-

Table 3 Average information criteria when data conformed to the 2-class Mix2PL model ($n=500$)

J	Candidate model	Model selection method										
		LOO	p_{LOO}	WAIC	p_{WAIC}	WBIC	DIC	p_{DIC}	AIC	BIC	CAIC	SABIC
15	2PL	7670.369	395.131	7650.475	385.184	10,326.803	7662.095	439.741	**7284.354**	**7415.007**	**7446.007**	**7316.611**
	2C-Mix2PL	**7618.833**	**343.533**	**7608.989**	**338.611**	10,237.826	**7605.658**	**362.867**	7368.791	7634.311	7697.311	7434.345
	3C-Mix2PL	7643.923	345.446	7635.301	341.135	**10,169.990**	7635.656	365.512	7457.348	7857.736	7952.736	7556.200
20	2PL	10,215.392	428.008	10,200.082	420.353	13,268.937	10,205.400	465.396	**9822.004**	**9994.803**	**10,035.803**	**9864.667**
	2C-Mix2PL	10,148.694	**373.051**	**10,139.896**	**368.651**	12,881.204	**10,124.379**	**379.086**	9911.293	10,261.106	10,344.106	9997.659
	3C-Mix2PL	**10,148.386**	374.426	10,140.550	370.508	**12,817.232**	10,129.247	384.534	9994.714	10,521.539	10,646.539	10,124.782
30	2PL	14,939.401	472.259	14,928.928	467.023	17,747.333	14,930.517	503.421	**14,549.097**	**14,806.188**	**14,867.188**	**14,612.570**
	2C-Mix2PL	14,856.803	**422.834**	14,848.819	**418.842**	**16,758.383**	**14,802.822**	**397.725**	14,651.096	15,169.492	15,292.492	14,779.083
	3C-Mix2PL	**14,855.829**	424.094	**14,848.657**	420.508	16,843.128	14,812.664	408.723	14,773.941	15,553.643	15,738.643	14,966.442

Values in bold represent the smallest values among the three candidate models for each information criteria and the associated effective number of parameters

Table 4 Average information criteria when data conformed to the 2-class Mix2PL model ($n = 1000$)

J	Candidate model	Model selection method										
		LOO	p_{LOO}	WAIC	p_{WAIC}	WBIC	DIC	p_{DIC}	AIC	BIC	CAIC	SABIC
15	2PL	15,163.495	789.313	15,126.598	770.865	20,669.410	15,134.948	863.819	**14,333.129**	**14,485.269**	**14,516.269**	**14,386.812**
	2C-Mix2PL	15,049.182	**672.035**	**15,032.260**	**663.575**	20,581.755	**15,021.678**	**704.788**	14,442.891	14,752.078	14,815.078	14,551.987
	3C-Mix2PL	**15,048.999**	673.961	15,033.887	666.405	**20,492.761**	15,033.999	717.391	14,506.608	14,972.846	15,067.846	14,671.120
20	2PL	19,945.572	838.982	19,917.181	824.787	26,706.559	19,917.495	902.349	**19,097.146**	**19,298.364**	**19,339.364**	**19,168.146**
	2C-Mix2PL	19,805.440	**719.506**	**19,790.239**	**711.906**	26,048.928	**19,757.881**	**728.066**	19,195.815	19,603.160	19,686.160	19,339.546
	3C-Mix2PL	**19,805.302**	721.681	19,791.810	714.936	**25,927.445**	19,778.452	749.098	19,279.354	19,892.824	20,017.824	19,495.817
30	2PL	30,201.644	905.524	30,183.281	896.343	36,100.450	30,176.203	954.231	**29,343.972**	**29,643.345**	**29,704.345**	**29,449.606**
	2C-Mix2PL	30,015.540	**794.212**	**30,002.250**	**787.566**	**34,327.773**	**29,922.449**	**752.610**	29,415.840	30,019.493	30,142.493	29,628.839
	3C-Mix2PL	**30,014.453**	796.654	30,002.467	790.661	34,504.287	29,955.879	787.864	29,538.014	30,445.949	30,630.949	29,858.379

Values in bold represent the smallest values among the three candidate models for each information criteria and the associated effective number of parameters

mation criteria (namely, AIC, BIC, CAIC, and SABIC) consistently preferred the simpler 2PL model regardless of test length or sample size.

- On the other hand, the three Bayesian information criteria, LOO, WAIC and DIC, tended to choose either the correct two-class model or a more complicated three-class model. Among the three information criteria, DIC performed the best, followed by WAIC, and then LOO.
- Performance of DIC or WAIC seemed to be affected by sample size, as larger sample sizes tended to result in a higher likelihood of selecting the correct model; for example, with $J=15$, DIC (WAIC) selected the correct two-class Mix2PL model 76% (68%) of the time when $n=500$ vs. 84% (72%) of the time when $n=1000$. This is, however, not observed with other information criteria.
- WBIC performed poorly when $J<30$ as it consistently selected the more complicated model; when $J=30$, it was able to identify the correct model more than 95% of the time regardless of sample size.

These findings are consistent to those from Tables 3 and 4, where we see that after averaging, AIC, BIC, CAIC, and SABIC were consistently smaller for the simpler 2PL model. DIC, on the other hand, consistently selected the correct two-class Mix2PL model. Among the three fully Bayesian information criteria, LOO tended to favor the more complex three-class model. It is, however, noted that the three-class solution did not differ much from the two-class solution as the difference of the average values of LOO between the two-class Mix2PL model (correct model) and the three-class Mix2PL model was less than 1.0 when $n=1000$ or $J>15$. As a matter of fact, average LOO and WAIC values for fitting the two-class and three-class models were almost identical for these test length and sample size conditions. In addition, the average WAIC was able to identify the correct two-class model across all the simulated conditions except when $n=500$ and $J=30$. The average WBIC, on the other hand, performed better when test length increased as it selected the correct model only when $J=30$ and in other conditions, WBIC preferred the more complicated three-class model. It is further noted that the average effective number of parameters for WAIC, LOO or DIC was the smallest for fitting the correct two-class Mix2PL model.

Furthermore, as suggested by Gelman et al. (2014), when deciding on the best fitting model, the effective number of parameters associated with Bayesian information criteria should also be taken into account, especially when the differences between the values of these measures for the candidate models are small, such that the simpler model is preferred over the more complex one. This is particularly true for LOO and WAIC, as the average effective number of parameters for the two measures as presented in Tables 3 and 4 indicates that the two-class Mix2PL model was the least complex and shall be preferred although LOO tended to favor the more complex three-class Mix2PL model across all the simulated conditions except when $n=500$ and $J=15$ and WAIC tended to favor the more complex three-class Mix2PL model when $n=500$ and $J=30$. For example, in conditions where $n=500$ and $J=30$, the average effective number of parameters for the two-class Mix2PL model ($p_{\text{LOO}}=422.834$, $p_{\text{WAIC}}=418.842$) was relatively smaller compared to the

Table 5 Frequency (relative frequency) for selecting each candidate model when data conformed to the 2PL model ($n = 500$)

J	Candidate model	Model selection method							
		LOO	WAIC	WBIC	DIC	AIC	BIC	CAIC	SABIC
15	2PL	**30 (0.60)**	**39 (0.78)**	**48 (0.96)**	**41 (0.82)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	12 (0.24)	6 (0.12)	2 (0.04)	5 (0.10)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	8 (0.16)	5 (0.10)	0 (0.00)	4 (0.08)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)
20	2PL	**37 (0.74)**	**44 (0.88)**	**47 (0.94)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	10 (0.20)	6 (0.12)	3 (0.06)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	3 (0.06)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)
30	2PL	**39 (0.78)**	**42 (0.84)**	**44 (0.88)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	6 (0.12)	4 (0.08)	3 (0.06)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	5 (0.10)	4 (0.08)	3 (0.06)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)

The maximum frequency of selecting a model is 50. Values in bold represent the largest frequency out of the 50 replications

three-class Mix2PL model ($p_{\text{LOO}} = 424.094$, $p_{\text{WAIC}} = 420.508$) or the 2PL IRT model ($p_{\text{LOO}} = 472.259$, $p_{\text{WAIC}} = 467.023$) although LOO or WAIC was slightly larger for the two-class model (LOO = 14,856.803, WAIC = 14,848.819) as compared to the three-class model (LOO = 14,855.829, WAIC = 14,848.657). Hence, we recommend that when using LOO or WAIC for comparing candidate models in practice, one shall also consider their effective number of parameters, especially when the difference between these values for candidate models is small. It is also noted that the conventional 2PL model (i.e., the one-class Mix2PL model), with relatively larger average values, was never preferred by any of the Bayesian information criteria when data involved multiple subpopulations.

5.2 Results for tests with a single population

The results for the second six conditions where data conformed to the conventional 2PL IRT model are summarized in Tables 5, 6, 7 and 8 with Tables 5 and 6 display the number and proportion of the time each model was identified as the best fitting model according to each information criterion, and Tables 7 and 8 show the average information criteria and effective number of parameters averaged across 50 replications, where the numbers in bold in Tables 5 and 6 are for the model with the highest frequency/proportion of being chosen as the best-fitting model, and those in Tables 7 and 8 are the smallest average values among the three candidate models. A close examination of Tables 5 and 6 indicate the following:

- The four frequentist information criteria, namely AIC, BIC, CAIC, and SABIC, consistently selected the correct 2PL model regardless of sample size or test length.

Table 6 Frequency (relative frequency) for selecting each candidate model when data conformed to the 2PL model ($n = 1000$)

J	Candidate model	Model selection method							
		LOO	WAIC	WBIC	DIC	AIC	BIC	CAIC	SABIC
15	2PL	**39 (0.78)**	**44 (0.88)**	**42 (0.84)**	**43 (0.86)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	8 (0.16)	5 (0.10)	8 (0.16)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	3 (0.06)	1 (0.02)	0 (0.00)	7 (0.14)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)
20	2PL	**31 (0.62)**	**39 (0.78)**	**44 (0.88)**	**47 (0.94)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	14 (0.28)	9 (0.18)	5 (0.10)	2 (0.04)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	5 (0.10)	2 (0.04)	1 (0.02)	1 (0.02)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)
30	2PL	**32 (0.64)**	**40 (0.80)**	**42 (0.84)**	**47 (0.94)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**	**50 (1.00)**
	2C-Mix2PL	8 (0.16)	5 (0.10)	4 (0.08)	2 (0.04)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	3C-Mix2PL	10 (0.20)	5 (0.10)	4 (0.08)	1 (0.02)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	Total	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)	50 (1.00)

The maximum frequency of selecting a model is 50. Values in bold represent the largest frequency out of the 50 replications

- DIC identified the correct 2PL model fairly well especially for $J \geq 20$, where its accuracy ranges from 94 to 100%.
- Among the three fully Bayesian information criteria, with a few replications favoring the two- or three-class MixIRT model, LOO and WAIC tended to go with more complex models with LOO being more toward this tendency; WBIC tended to favor the correct 2PL model more frequently and hence is more preferred than WAIC or LOO.

These findings are consistent with results presented in Tables 7 and 8. Specifically, after averaging, LOO, WAIC, WBIC, AIC, BIC, CAIC, and SABIC consistently favored the correct 2PL model regardless of sample size or test length. Test length may have an effect on the average DIC, as it tended to favor a more complex MixIRT model when $J = 15$. It is noted that some of the average LOO values for the two-class solution did not differ much from the one-class solution as the difference of the average values of LOO between the 2PL model (correct model) and the two-class Mix2PL model was only 0.18 for $n = 500$ and $J = 15$. In effect, the Bayesian approach resulted in almost identical average information (especially LOO and WAIC) in fitting the correct 2PL model and the more complicated MixIRT models regardless of sample size or test length. Moreover, the average effective number of parameters for LOO or WAIC is consistently the smallest for the correct 2PL model.

The average effective number of parameters associated with both LOO and WAIC, as presented in Tables 7 and 8, indicates that the 2PL model was the least complex across all the conditions. For example, in conditions where $n = 1000$ and $J = 15$, the average effective number of parameters for the 2PL IRT model ($p_{\text{LOO}} = 722.034$, $p_{\text{WAIC}} = 711.098$) was relatively smaller compared to that of the

Table 7 Average information criteria when data conformed to the 2PL model ($n = 500$)

J	Candidate model	Model selection method										
		LOO	p_{LOO}	WAIC	p_{WAIC}	WBIC	DIC	p_{DIC}	AIC	BIC	CAIC	SABIC
15	2PL	**8218.198**	**371.657**	**8205.657**	**365.386**	**9173.410**	8190.992	**385.581**	**7867.411**	**7998.064**	**8029.064**	**7899.668**
	2C-Mix2PL	8218.378	375.577	8207.142	369.959	9186.988	**8190.048**	387.119	7928.928	8194.449	8257.449	7994.483
	3C-Mix2PL	8219.455	377.448	8208.928	372.184	9206.592	8190.912	388.066	7992.846	8393.234	8488.234	8091.698
20	2PL	**10,905.548**	**409.585**	**10,894.515**	**404.069**	**12,258.972**	**10,881.899**	**425.388**	**10,538.512**	**10,711.311**	**10,752.311**	**10,581.174**
	2C-Mix2PL	10,906.927	413.398	10,897.022	408.445	12,277.846	10,888.798	433.512	10,621.286	10,971.098	11,054.098	10,707.651
	3C-Mix2PL	10,908.345	415.534	10,899.006	410.865	12,301.431	10,892.972	437.804	10,705.169	11,231.994	11,356.994	10,835.237
30	2PL	**16,190.934**	**457.303**	**16,183.378**	**453.525**	**18,017.056**	**16,167.282**	**467.018**	**15,822.264**	**16,079.355**	**16,140.355**	**15,885.738**
	2C-Mix2PL	16,192.618	460.524	16,185.660	457.045	18,038.848	16,172.970	473.453	15,945.517	16,463.914	16,586.914	16,073.504
	3C-Mix2PL	16,193.921	462.230	16,187.311	458.926	18,058.034	16,176.763	477.171	16,069.592	16,849.294	17,034.294	16,262.093

Values in bold represent the smallest values among the three candidate models for each information criteria and the associated effective number of parameters

Table 8 Average information criteria when data conformed to the 2PL model ($n = 1000$)

J	Candidate model	Model selection method										
		LOO	p_{LOO}	WAIC	p_{WAIC}	WBIC	DIC	p_{DIC}	AIC	BIC	CAIC	SABIC
15	2PL	**16,243.005**	**722.034**	**16,221.133**	**711.098**	**18,137.258**	16,200.562	758.563	**15,503.999**	**15,656.139**	**15,687.139**	**15,557.683**
	2C-Mix2PL	16,244.445	728.422	16,224.397	718.398	18,149.129	16,207.768	768.986	15,564.782	15,873.970	15,936.970	15,673.879
	3C-Mix2PL	16,245.978	732.171	16,226.936	722.650	18,175.923	**16,171.234**	**733.650**	15,627.584	16,093.822	16,188.822	15,792.095
20	2PL	**21,900.728**	**780.980**	**21,884.032**	**772.632**	**24,421.594**	**21,862.890**	**813.020**	**21,131.870**	**21,333.088**	**21,374.088**	**21,202.870**
	2C-Mix2PL	21,901.210	787.560	21,885.775	779.842	24,439.814	21,870.031	824.762	21,211.269	21,618.613	21,701.613	21,355.000
	3C-Mix2PL	21,903.347	791.056	21,888.745	783.756	24,469.596	21,874.371	829.556	21,294.815	21,908.285	22,033.285	21,511.277
30	2PL	**31,941.349**	**872.054**	**31,929.708**	**866.234**	**35,836.107**	**31,908.578**	**899.200**	**31,131.378**	**31,430.753**	**31,491.753**	**31,237.013**
	2C-Mix2PL	31,942.514	877.082	31,931.416	871.533	35,870.744	31,914.724	908.344	31,252.381	31,856.035	31,979.035	31,465.380
	3C-Mix2PL	31,944.200	879.936	31,933.549	874.611	35,886.802	31,919.305	913.537	31,375.768	32,283.704	32,468.704	31,696.133

Values in bold represent the smallest values among the three candidate models for each information criteria and the associated effective number of parameters

two-class Mix2PL model ($p_{\mathrm{LOO}}=728.422$, $p_{\mathrm{WAIC}}=718.398$) or the three-class Mix2PL model ($p_{\mathrm{LOO}}=732.171$, $p_{\mathrm{WAIC}}=722.650$). Therefore, following our recommendation for tests involving multiple latent subpopulations, we again see the potential benefit of considering the effective number of parameters when LOO or WAIC is used in comparing candidate models where a single latent population is assumed.

6 Discussion and conclusion

This study focuses on the performance of fully Bayesian information criteria, namely, LOO, WAIC, and WBIC, in terms of the accuracy in determining the number of latent classes of the Mix2PL model while comparing it to the conventional 2PL model and further in terms of comparison with other information criteria including AIC, BIC, CAIC, SABIC, and DIC.

Regarding the accuracy in determining the number of latent classes, for the condition where data conformed to the two-class Mix2PL model, the results indicate that among the three fully Bayesian information criteria, WBIC shall only be used when tests consist of at least 30 items; WAIC performed slightly better than LOO in recovering the number of latent classes, although the proportion of the time the correct model was selected as the best fitting model, for both measures, decreased compared to the situation where the generated model was the conventional 2PL model. When considering other information criteria, it is found that the frequentist information criteria all failed to identify the correct model as they consistently favored the simpler 2PL IRT model. DIC, on the other hand, has outperformed both the frequentist and fully-Bayesian information criteria. On the other hand, for the condition where data conformed to the conventional 2PL model, the results indicate that among the three fully Bayesian information criteria, WBIC performed better than WAIC or LOO alone in recovering the number of latent classes in terms of the proportion of the time the correct model was selected as the best fitting model. In addition, when considering other information criteria, it is found that all the frequentist information criteria consistently favored the correct 2PL model and the Bayesian information criterion DIC performed similarly if not better than WBIC. Summarizing across the two simulated test conditions, we make the following conclusions:

1. AIC, BIC, and their related CAIC and SABIC tend to select the simpler model and are not recommended when the actual data involve multiple latent classes. While this finding differs from those from previous studies evaluating their performances via other fully Bayesian estimation algorithms such as Gibbs sampling, it can be explained by citing Watanabe's (2021) argument that if AIC and BIC are used in model selection, the best model with the smallest values of both the generalization loss and the free energy are not chosen, rather tighter (or smaller) models are selected.
2. DIC alone performs equally well and sometimes better than some of the frequentist or fully-Bayesian criterion measures especially when sample size is relatively large (with e.g., 1000 or more subjects). This finding, however, differs from those from previous studies such as Li et al. (2009) and needs further validation.

Table 9 The average proportion of persons in each class for incorrect the 3-class Mix2PL model where the generating model was the 2-class Mix2PL model

		3-class Mix2PL		
		$\widehat{\pi}_1$	$\widehat{\pi}_2$	$\widehat{\pi}_3$
$J=15$	500	0.14	0.41	0.45
	1000	0.12	0.42	0.46
$J=20$	500	0.13	0.42	0.46
	1000	0.10	0.43	0.46
$J=30$	500	0.11	0.43	0.46
	1000	0.07	0.45	0.48

3. For the three fully Bayesian information criteria, WBIC can be used for detecting the number of latent classes for tests with at least 30 items; WAIC and LOO alone are not suggested for fitting the MixIRT models, and they shall be used together with their effective number of parameters in choosing the correct number of latent classes.

It is noted the fully Bayesian information criteria alone can sometimes favor a more complicated model such as the two- or three-class Mix2PL model when data involve one or two latent subpopulations, respectively. This can be explained by what Watanabe (2021) noted: for a normal mixture model such as MixIRT, neither the generalization loss estimated by WAIC (or LOO) nor the free energy estimated by WBIC increases "even if the statistical model is redundant" (p. 18) as the case with regular statistical models; thus, WAIC (and similarly LOO) or WBIC does not increase either and hence the more complicated (or redundant) model may be selected more often.

In addition, for test conditions where data conformed to the two-class Mix2PL model, it is noted that although in the simulation study, LOO, WAIC, and WBIC sometimes selected the more complex three-class Mix2PL model as the best fitting model, the average proportion of persons (across the 50 replications) for one of the three classes was relatively low (see Table 9). For example, for test situations with 15 items, the average proportion of persons in one of the three classes was 0.14 with a sample size of 500 and 0.12 with a sample size of 1000. Same observation was made for the condition where data conformed to the conventional 2PL model, where LOO, WAIC, and WBIC sometimes selected the more complex and yet incorrect two-class Mix2PL model as the best fitting model. The average proportion of persons (across the 50 replications) for each class is summarized in Table 10 and suggests that the average proportion for one of the two classes was relatively very low. For example, for tests with 20 items, the average proportion of persons in one of the two classes of the Mix2PL model was 0.03 for both sample sizes, 500 and 1000. These results, in general, indicate that when a more complex while incorrect model was selected by one or more of the fully Bayesian criteria without considering the effective number of parameters, the proportion of persons in one of the classes can be relatively low.

Table 10 The average proportion of persons in each class for the incorrect 2-class and 3-class Mix2PL models where the generating model was the conventional 2PL model

		2-class Mix2PL		3-class Mix2PL		
		$\widehat{\pi}_1$	$\widehat{\pi}_2$	$\widehat{\pi}_1$	$\widehat{\pi}_2$	$\widehat{\pi}_3$
$J=15$	500	0.06	0.94	0.03	0.05	0.92
	1000	0.03	0.97	0.02	0.04	0.94
$J=20$	500	0.03	0.97	0.02	0.03	0.95
	1000	0.03	0.97	0.02	0.03	0.96
$J=30$	500	0.02	0.98	0.01	0.02	0.97
	1000	0.02	0.98	0.01	0.02	0.97

Given the results displayed in Tables 9 and 10, caution should be taken when fully Bayesian criteria such as LOO, WAIC, or WBIC select a more complicated model as the best fitting model. A further step should be taken to understand the nature of the estimated mixing proportions of the complex model favored by such information criteria.

Our findings that LOO tends to favor a more complex model and is less preferred than WAIC for fitting MixIRT models, however, differ from those of Luo and Al-Harbi (2017), who compared WAIC and LOO for conventional IRT models and concluded that WAIC had a slightly lower detection rate than LOO (although the difference is negligible) in the condition where the generating model was the conventional one-parameter IRT model. Similarly, these results differ from those of Luo (2019) who compared WAIC and LOO with the DIC, AIC, BIC, AICc, SABIC for polytomous IRT models, where the results indicated the detection rate of WAIC (0.935) was slightly lower than that of LOO (0.946). This difference may be a result of the models considered; namely, LOO tends to work well for unidimensional models with varying number of parameters whereas WAIC tends to work well for models with more complex latent structure. This certainly needs to be confirmed with additional studies.

Further, our result where the frequentist information criteria consistently favored the simpler IRT model when the test involved multiple latent subpopulations could be directly tied with how these indices were computed in the fully Bayesian framework. In this study, we followed Congdon (2003) to replace the MLE-based deviance, $-2\log(L)$, with the posterior mean of the deviance as it has been commonly adopted in studies on MixIRT models. While this result differed from previous studies that warrant additional investigations, it suggests potential limitations of using the posterior mean of the deviance as suggested by Congdon (2003) for fully Bayesian estimation especially with non-random walk MCMC algorithms, as it may not correspond to the maximized likelihood, and hence calls for alternative approaches. Additional research should be carried out to further evaluate this.

Given the computational expense, we focused on the two-parameter models in this study to consider a few sample-size and test-length conditions assuming equal mixing proportions for test situations where data involve one or two subpopulations. Future research can consider other test conditions or those with more than two latent classes. Additional studies can also investigate the performance of these fully

Bayesian information criteria in selecting the true model using different MixIRT models such as the dichotomous mixture one-parameter (Mix1PL) or three-parameter (Mix3PL) model or polytomous MixIRT models. In addition, for simulation conditions where data involved two latent classes, person ability parameters were generated from distributions with their locations being four standard deviations apart, i.e., $\theta_1 \sim N(-2, 1)$ and $\theta_2 \sim N(2, 1)$, to minimize potential overlap between the two subpopulations. Further studies can consider situations where the latent classes are from distributions with closer location parameters. Finally, future study can consider implementing Mixture IRT models using NUTS to real data problems to identify the number of latent classes using the fully Bayesian information criteria, while following guidelines from the Monte Carlo simulations.

Appendix 1. Stan code for computing the posterior distribution under the inverse temperature $\beta = 1/\log(n)$

```
data {
  int<lower=1>K;
  int<lower=1>N;
  int<lower=1>J;
  int<lower=0,upper=1>y[N,J];
  vector<lower=0>[K] dir_alpha;}

parameters {
  simplex[K] pi;
  ordered[K] mu;
  ordered[K] beta[J];
  vector<lower=0>[K] alpha[J];
  vector[N] theta;}

model {
  real lps[K];
  real p[K];
  real lpth[K];
  for (k in 1:K){
  mu[k]~normal(0, 1);}
  pi~dirichlet(dir_alpha);
  for (j in 1:J){
  for (k in 1:K){
  beta[j,k]~normal (0,1);
  alpha[j,k]~normal(0,1);}}
  for (i in 1:N){
  for (k in 1:K){
  lpth[k]=log(pi[k])+normal_lpdf(theta[i]|mu[k], 1);}
  target+=log_sum_exp(lpth);}
```

```
    for (i in 1:N){
    for (j in 1:J){
    for (k in 1:K){
    p[k] = inv_logit( alpha[j,k]*(theta[i]-beta[j,k]));
    lps[k] = log(pi[k]) + bernoulli_lpmf(y[i,j] | p[k]);}
    target + = (1.0/log(N)) * log_sum_exp(lps);}}}
    generated quantities{
    real p[K];
    vector[K] log_likk[N,J];
    vector[J] log_lik[N];
    for (i in 1:N){
    for (j in 1:J){
    for (k in 1:K){
    p[k] = inv_logit( alpha[j,k]*(theta[i]-beta[j,k]));
    log_likk[i, j, k] = log(pi[k]) + bernoulli_lpmf(y[i,j] | p[k]);}
    log_lik[i, j] = log_sum_exp(log_likk[i, j]);}}}
```

Appendix 2. Stan code for computing the posterior distribution under the inverse temperature $\beta = 1$

```
data {
    int < lower = 1 > K;
    int < lower = 1 > N;
    int < lower = 1 > J;
    int < lower = 0,upper = 1 > y[N,J];
    vector < lower = 0 > [K] dir_alpha;}

parameters {
    simplex[K] pi;
    ordered[K] mu;
    ordered[K] beta[J];
    vector < lower = 0 > [K] alpha[J];
    vector[N] theta;}

model {
    real lps[K];
    real p[K];
    real lpth[K];
    for (k in 1:K){
    mu[k] ~ normal(0, 1);}
    pi ~ dirichlet(dir_alpha);
    for (j in 1:J){
    for (k in 1:K){
    beta[j,k] ~ normal (0,1);
```

```
alpha[j,k] ~ normal(0,1);}}
for (i in 1:N){
for (k in 1:K){
lpth[k] = log(pi[k]) + normal_lpdf(theta[i] | mu[k], 1);}
target + = log_sum_exp(lpth);}
for (i in 1:N){
for (j in 1:J){
for (k in 1:K){
p[k] = inv_logit( alpha[j,k]*(theta[i]-beta[j,k]));
lps[k] = log(pi[k]) + bernoulli_lpmf(y[i,j] | p[k]);}
target + = log_sum_exp(lps);}}}
generated quantities{
real p[K];
vector[K] log_likk[N,J];
vector[J] log_lik[N];
for (i in 1:N){
for (j in 1:J){
for (k in 1:K){
p[k] = inv_logit( alpha[j,k]*(theta[i]-beta[j,k]));
log_likk[i, j, k] = log(pi[k]) + bernoulli_lpmf(y[i,j] | p[k]);}
log_lik[i, j] = log_sum_exp(log_likk[i, j]);}}}
```

Acknowledgements We thank the editor and two anonymous reviewers for their valuable comments, and the University of Chicago's Research Computing Center for their support of this work.

Funding The authors have not disclosed any funding.

Declarations

Conflict of interest There are no financial conflicts of interest to disclose.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723

Al Hakmani R, Sheng Y (2019) NUTS for mixture IRT models. In: Wiberg M, Culpepper S, Janssen R, González J, Molenaar D (eds) Quantitative psychology. Springer, New York, pp 25–37

Andrich D (1978) A rating formulation for ordered response categories. Psychometrika 43(4):561–573

Bilir MK (2009) Mixture item response theory-MIMIC model: simultaneous estimation of differential item functioning for manifest groups and latent classes. Doctoral dissertation. ProQuest Dissertations & Theses A&I. (Order No. 3399179)
Birnbaum A (1968) Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR (eds) Statistical theories of mental test scores. Addison-Wesley, Reading, pp 397–479
Bock RD (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37:29–51
Bolt DM, Cohen AS, Wollack JA (2001) A mixture item response model for multiple-choice data. J Educ Behav Stat 26(4):381–409
Bolt DM, Cohen AS, Wollack JA (2002) Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. J Educ Meas 39(4):331–348
Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52(3):345–370
Brooks S, Smith J, Vehtari A, Plummer M, Stone M, Robert CP et al (2002) Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde. J R Stat Soc Ser B Stat Methodol 64:616–639
Carlin BP, Louis TA (2001) Bayes and empirical Bayes methods for data analysis, 2nd edn. Chapman & Hall/CRC, Boca Raton
Cho S-J, Cohen AS, Kim S-H (2013) Markov chain Monte Carlo estimation of a mixture item response theory model. J Stat Comput Simul 83:278–306. https://doi.org/10.1080/00949655.2011.603090
Choi IH, Paek I, Cho SJ (2017) The impact of various class-distinction features on model selection in the mixture Rasch model. J Exp Educ 85(3):411–424. https://doi.org/10.1080/00220973.2016.1250208
Congdon P (2003) Applied Bayesian modelling. Wiley, New York
Da Silva MA, Bazán JL, Huggins-Manley AC (2018) Sensitivity analysis and choosing between alternative polytomous IRT models using Bayesian model comparison criteria. Commun Stat Simul Comput 48:601–620. https://doi.org/10.1080/03610918.2017.1390126
De Ayala RJ, Kim SH, Stapleton LM, Dayton CM (2002) Differential item functioning: a mixture distribution conceptualization. Int J Test 2(3&4):243–276
de la Torre J, Stark S, Chernyshenko OS (2006) Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. Appl Psychol Meas 30(3):216–232. https://doi.org/10.1177/0146621605282772
Duane S, Kennedy A, Pendleton BJ, Roweth D (1987) Hybrid Monte Carlo. Phys Lett B 195:216–222. https://doi.org/10.1016/0370-2693(87)91197-X
Finch WH, French BF (2012) Parameter estimation with mixture item response theory models: a Monte Carlo comparison of maximum likelihood and Bayesian methods. J Mod Appl Stat Methods 11(1):167–178
Fisher (1922) On the mathematical foundation of theoretical Statistics. Philos Trans R Soc 222:309–368
Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–472
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014) Bayesian data analysis, 3rd edn. Chapman & Hall/CRC, Boca Raton, FL
Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6(6):721–741. https://doi.org/10.1109/TPAMI.1984.4767596
Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97–109. https://doi.org/10.1093/biomet/57.1.97
Hoffman MD, Gelman A (2014) The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 15(2):1593–1624
Jang Y, Cohen AS (2020) The impact of Markov chain convergence on estimation of mixture IRT model parameters. Educ Psychol Meas 80(5):975–994. https://doi.org/10.1177/0013164419898228
Lee H, Beretvas SN (2014) Evaluation of two types of differential item functioning in factor mixture models with binary outcomes. Educ Psychol Meas 74(5):831–858. https://doi.org/10.1177/0013164414526881

Li F, Cohen A, Kim S, Cho S (2009) Model selection methods for mixture dichotomous IRT models. Appl Psychol Meas 33(5):353–373. https://doi.org/10.1177/0146621608326422
Luo Y (2019) LOO and WAIC as model selection methods for polytomous items. Psychol Test Assess Model 61:161–185
Luo Y, Al-Harbi K (2017) Performances of LOO and WAIC as IRT model selection methods. Psychol Test Assess Model 59(2):183–205
Luo Y, Jiao H (2017) Using the Stan program for Bayesian item response theory. Educ Psychol Meas 78(3):384–408
Masters GN (1982) A Rasch model for partial credit scoring. Psychometrika 47(2):149–174
Metropolis N, Ulam S (1949) The Monte Carlo method. J Am Stat Assoc 44(247):335–341
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092
Meyer JP (2010) A mixture Rasch model with Item response time components. Appl Psychol Meas 34(7):521–538. https://doi.org/10.1177/0146621609355451
Muraki E (1992) A generalized partial credit model: application of an EM algorithm. Appl Psychol Meas 16(2):159–176
Neal RM (2011) MCMC using Hamiltonian dynamics. In: Brooks S, Gelman A, Jones G, Meng X (eds) Handbook of Markov chain Monte Carlo. CRC Press, Boca Raton, pp 113–162
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans A Math Phys Eng Sci 231:289–337. https://doi.org/10.1098/rsta.1933.0009
Nylund KL, Asparouhov T, Muthén BO (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Struct Equ Model 14:535–569. https://doi.org/10.1080/10705510701575396
Plummer M (2008) Penalized loss functions for Bayesian model comparison. Biostatistics 9:523–539. https://doi.org/10.1093/biostatistics/kxm049
Preinerstorfer D, Formann AK (2012) Parameter recovery and model selection in mixed Rasch models. Br J Math Stat Psychol 65(2):251–262. https://doi.org/10.1111/j.2044-8317.2011.02020.x
Rost J (1990) Rasch models in latent classes: an integration of two approaches to item analysis. Appl Psychol Meas 14(3):271–282. https://doi.org/10.1177/014662169001400305
Samejima F (1969) Estimation of latent ability using a response pattern of graded scores. Psychometrika 17:1–37
Samuelsen K (2005) Examining differential item functioning from a latent class perspective. Doctoral dissertation. ProQuest Dissertations & Theses A&I. (Order No. 3175148)
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Sclove SL (1987) Application of model-selection criteria to some problems in multivariate analysis. Psychometrika 52(3):333–343. https://doi.org/10.1007/BF02294360
Sen S, Cohen AS, Kim SH (2016) The impact of non-normality on extraction of spurious latent classes in mixture IRT models. Appl Psychol Meas 40(2):98–113. https://doi.org/10.1177/0146621615605080
Sen S, Cohen AS, Kim S (2019) Model selection for multilevel mixture Rasch models. Appl Psychol Meas 43(4):272–289. https://doi.org/10.1177/0146621618779990
Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B Stat Methodol 64(4):583–639
Stan Development Team (2020) RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.
Sugiura N (1978) Further analysts of the data by Akaike's information criterion and the finite corrections: further analysts of the data by Akaike's. Commun Stat Theory Methods 7(1):13–26
Uto M, Ueno M (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika 47:469–496
Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27(5):1413–1432. https://doi.org/10.1007/s11222-016-9696-4
Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 11:3571–3594
Watanabe S (2013) A widely applicable Bayesian information criterion. J Mach Learn Res 14:867–897
Watanabe S (2021) WAIC and WBIC for mixture models. Behaviormetrika 48:5–21

Wollack JA, Bolt DM, Cohen AS, Lee YS (2002) Recovery of item parameters in the nominal response model: a comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. Appl Psychol Meas 26(3):339–352. https://doi.org/10.1177/0146621602026003007