Bayesian Psychometrics

Xiang Liu, Matthew S. Johnson, and Zhuangzhuang Han

Educational Testing Service

Abstract

Bayesian statistical methods have become popular in the field of psychometrics. It provides a class of approaches alternative to classical or frequentist statistical statistics (e.g. maximum likelihood, confidence intervals, p-values). In certain cases, Bayesian approaches can alleviate some difficulties classical approaches may have. In this chapter, using a 1-parameter normal ogive IRT model as an example, we introduce important Bayesian statistics concepts such as model formulation, posterior inference, MCMC, and posterior predictive model checks. An empirical example is provided for demonstration purpose.

Keywords: Bayesian, psychometrics, IRT, MCMC, posterior predictive model checks

Bayesian Psychometrics

Introduction

Bayesian statistical methods have become popular in the field of psychometrics. Thanks to the development of general purpose Bayesian inference software (e.g *Stan*; Gelman et al., 2015), researchers are now able to quickly develop and fit sophisticated and complex psychometric models that once require significant computational expertise to estimate. In addition, Bayesian methods also provide a class of approaches alternative to classical or frequentist statistical statistics (e.g. maximum likelihood, confidence intervals, p-values). In certain cases, Bayesian approaches can alleviate some difficulties classical approaches may have. In this chapter, using a 1-parameter normal ogive IRT model as an example, we introduce important Bayesian statistics concepts such as model formulation, posterior inference, MCMC, and posterior predictive model checks. An empirical example is provided for demonstration purpose.

The 1-parameter normal ogive IRT model

Let $X_{ij} = x_{ij}$ be a dichotomous random variable of the item response from the i^{th} student to the j^{th} item, $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$. And $x_{ij} = 1$ if the response is correct, $x_{ij} = 0$ otherwise. Under the 1-parameter normal ogive (1PNO) IRT model, the probability of a correct response is given by

$$P(X_{ij} = 1|\theta_i, a, b_j) = \Phi(a(\theta_i - b_j)), \tag{1}$$

where $\Phi(\cdot)$ denotes the cumulative probability density function (CDF) of the standard normal distribution, θ_i is the person parameter, b_j is the item difficulty parameter, and a is the common discrimination parameter shared among J items. The person parameter is assumed to follow a standard normal distribution, i.e. $\theta_i \sim N(0, 1)$. Because a is a constant across items, the item characteristic curves (ICC) never cross each other and only differ by locations (see Figure 1).

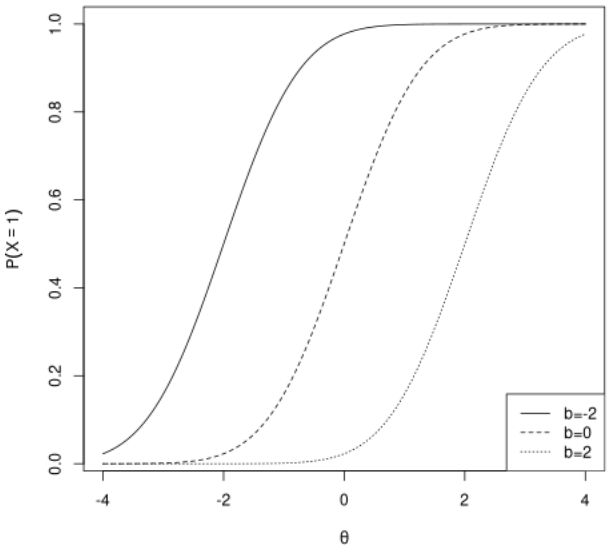

Figure 1

ICC of the 1PNO IRT model

1PNO as a Bayesian hierarchical model

The 1PNO as in Equation 1 can be expressed as a Bayesian hierarchical model. For computational efficiency, we reparameterize the model so that the model is in its slope-intercept form,

$$P(X_{ij} = 1|\theta_i, \psi_j) = \Phi(\theta_i + \psi_j), \tag{2}$$

and $\theta_i \sim N(0, a^2)$, where a^2 is the variance of the normal distribution. In Bayesian statistics, model parameters are treated as random and are given prior distributions. This is in contrast to the classical likelihood methods where model parameters are treated as fixed but unknown constants. A prior distribution describes any prior belief we may have (or the lack of it) on the parameter with uncertainty.

For instance, in the current example, there may be reasonable prior information on the item parameter ψ_j from similar items previously administered to a comparable population of students. In this case, we may choose a normal prior with mean μ_{ψ_j} informed by prior information and variance $\sigma^2_{\psi_j}$ expressing the level of uncertainty, i.e.

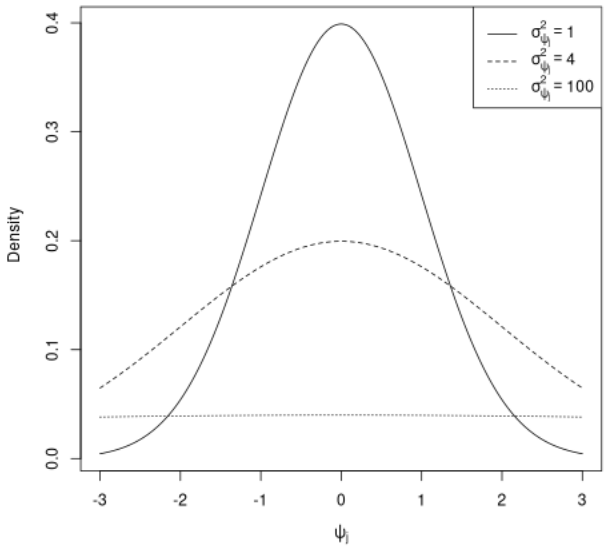

Figure 2

Density of prior distribution of ψ_j

$\psi_j \sim N(\mu_{\psi_j}, \sigma^2_{\psi_j})$. A prior with larger variance (or a flatter prior, see Figure 2) is less informative, thus having less impact on the posterior inference of the parameter. However, prior information may not always be available. In this case, instead of choosing a fixed prior with large variance, an alternative approach is to use a hierarchical prior. For each item parameter ψ_j, we assign a same Gaussian prior, $\psi_j \sim N(\mu_\psi, \sigma^2_\psi)$. Hyper-priors are assigned to the parameters of this prior distribution, i.e. $\mu_\psi \sim N(0, 10^2)$ and $\sigma^2_\psi \sim Ga^{-1}(10^{-4}, 10^{-4})$, where $Ga^{-1}(\cdot)$ denotes the inverse-gamma distribution. The chosen inverse gamma is noninformative if most of the posterior density is away from 0 as the prior density is flat in the tail. The choice of this set of priors is also mathematically convenient. The normal prior and inverse-gamma prior are conditionally conjugate for the mean parameter and the variance parameter of a normal model. We choose a similar noninformative inverse-gamma distribution for a^2, i.e. $a^2 \sim Ga^{-1}(10^{-4}, 10^{-4})$. We can then

write the 1PNO IRT model has a Bayesian hierarchical model,

$$
\begin{aligned}
x_{ij}|\theta_i, \psi_j &\sim Bernoulli(\Phi(\theta_i + \psi_j)) \\
\psi_j|\mu_\psi, \sigma^2_\psi &\sim N(\mu_\psi, \sigma^2_\psi) \\
\mu_\psi &\sim N(0, 10^2) \\
\sigma^2_\psi &\sim Ga^{-1}(10^{-4}, 10^{-4}) \\
\theta_i|a^2 &\sim N(0, a^2) \\
a^2 &\sim Ga^{-1}(10^{-4}, 10^{-4}).
\end{aligned} \tag{3}
$$

Gibbs sampling with augmented latent variables

For the estimation of the model, a direct Gibbs sampling is possible. However, we would need to augment latent variables which is similar to the typical Bayesian posterior computation approach for the probit regression models. For every item response x_{ij}, there is a latent variable y_{ij} associated. The relationship between them is deterministic such that

$$
x_{ij} = \begin{cases} 1 \text{ if } y_{ij} > 0, \\ 0 \text{ otherwise,} \end{cases}
$$

where $y_{ij} \sim N(\theta_i + \psi_j, 1)$. It is straightforward to verify that, under this specification, the probability of a correct response remains $\int_{-\infty}^{\theta_i+\psi_j} \phi(y)dy = \Phi(\theta_i + \psi_j)$.

A Gibbs sampler (Geman & Geman, 1984) works by iteratively sampling from full conditional distributions of the parameters (for a review, see Neal, 1998). The algorithm has been widely used in Bayesian estimation of various psychometric models (e.g. Culpepper, 2015; Fox and Glas, 2001; Liu and Johnson, 2019; Maris et al., 2015). To derive the densities full conditional distributions, we use the definition of conditional probabilities and the chain rule. For example,

$$
p(y_{ij}|x_{ij}, \theta_i, \psi_j) = \frac{p(x_{ij}|y_{ij})p(y_{ij}|\theta_i, \psi_j)}{\sum_{y_{ij}} p(x_{ij}|y_{ij})p(y_{ij}|\theta_i, \psi_j)}. \tag{4}
$$

Notice that the denominator of Equation 4 is a constant (with respect to y_{ij}) that

normalizes the full conditional density. Therefore, we can conveniently write

$$\begin{aligned} p(y_{ij}|x_{ij},\theta_i,\psi_j) &= Cp(x_{ij}|y_{ij})p(y_{ij}|\theta_i,\psi_j) \\ &\propto p(x_{ij}|y_{ij})p(y_{ij}|\theta_i,\psi_j). \end{aligned}$$

Given that $p(x_{ij}|y_{ij})$ is a piecewise point density with support on either $x_{ij}=1$ or $x_{ij}=0$ depending on y_{ij}, the full conditional distribution of y_{ij} is a truncated normal distribution, i.e.

$$y_{ij}|x_{ij},\theta_i,\psi_j \sim \begin{cases} N(\theta_i+\psi_j,1)T(0,) \text{ if } x_{ij}=1, \\ N(\theta_i+\psi_j,1)T(,0) \text{ if } x_{ij}=0. \end{cases}$$

$T(0,)$ denotes that the distribution is truncated with a lower bound of 0; while $T(,0)$ denotes a truncation with a upper bound of 0.

The full conditional distributions of other parameters can be derived similarly. Since we choose priors that are *conditionally* conjugate, the full conditional distributions of the parameters are within the same distribution families as their priors. The standard conjugacy results apply (for a reference, see Gelman et al., 2013). In the following section, we state the results without detailed derivations.

Now we can summarize the Gibbs sampler. After assigning initial values to parameters, at the T^{th} iteration,

1. for each i and j, update y_{ij} by drawing from $N(\theta_i+\psi_j,1)T(0,)$ if $x_{ij}=1$; else draw from $N(\theta_i+\psi_j,1)T(,0)$;
2. for each i, update θ_i by drawing from $N\left(\frac{1}{1/a^2+J}(\sum_j(y_{ij}-\psi_j)),(1/a^2+J)^{-1}\right)$;
3. for each j, update ψ_j by drawing from $N\left(\frac{1}{1/\sigma_\psi^2+N}(\mu_\psi/\sigma_\psi^2+\sum_i(y_{ij}-\theta_i)),\frac{1}{1/\sigma_\psi^2+N}\right)$;
4. update a^2 by drawing from $Ga^{-1}(10^{-4}+\frac{N}{2},10^{-4}+\frac{\sum_i\theta_i^2}{2})$;
5. update μ_ψ by drawing from $N\left(\frac{1}{1/10^2+J/\sigma_\psi^2}(\frac{\sum_j\psi_j}{\sigma_\psi^2}),(1/10^2+J/\sigma_\psi^2)^{-1}\right)$;
6. update σ_ψ^2 by drawing from $Ga^{-1}\left(10^{-4}+J/2,10^{-4}+\frac{\sum_j(\psi_j-\mu_\psi)^2}{2}\right)$.

Keep updating long enough, the draws are eventually random samples from the joint posterior distribution.

A real data example

We use the LSAT dataset provided by the *ltm R* package (Rizopoulos, 2015) for demonstration. The dataset was first described in Bock and Lieberman (1970). It contains dichotomous item responses from 1000 students to 5 items that are designed to measure a single latent trait. We fit the Bayesian 1PNO model with the Gibbs sampler described in the previous section. The posterior samples are then transformed back to the more common parameterization as in Equation 1. The posterior summary are in Table 1. The

Table 1

Posterior summary

	Mean	SD	lower	upper
b_1	-3.626	0.337	-4.299	-3.049
b_2	-1.405	0.157	-1.716	-1.128
b_3	-0.349	0.107	-0.569	-0.146
b_4	-1.823	0.188	-2.202	-1.500
b_5	-2.857	0.270	-3.405	-2.368
a	0.432	0.040	0.349	0.504

first two columns are posterior means and posterior standard deviations, which are perhaps most common in describing the posterior distributions. The last two columns are lower limit and upper limit of the 95% highest posterior density (HPD) intervals.

Posterior predictive model checking

A statistical model is essentially a set of assumptions. Before any inferences are drawn from the model, it is important to assess the fit of the model or, in other words, to check the degree to which the model fails to explain the data that are of our practical interest. In a Bayesian framework, this is typically done through posterior predictive model checking (PPMC; Gelman et al., 2013; Rubin, 1984). The basic intuition of the PPMC method is that we can check the observed data against the replicated data predicted by the model using some statistics. The distribution of the replicated data is also referred to as the posterior predictive distribution, i.e.

$$p(\boldsymbol{x}^{rep}|\boldsymbol{x}) = \int p(\boldsymbol{x}^{rep}|\boldsymbol{\zeta})p(\boldsymbol{\zeta}|\boldsymbol{x})d\boldsymbol{\zeta}, \tag{5}$$

where $\boldsymbol{\zeta}$ is a vector of all model parameters and $p(\boldsymbol{\zeta}|\boldsymbol{x})$ is the posterior distribution. Except for some simple models, finding the posterior predictive distribution requires sampling in general.

A statistic is a function of the data. Different statistics may describe different aspects of the data. Thus, it is important to choose appropriate statistics that reflect aspects of the model that are of interest to us. In this chapter, we follow the suggestions given by Sinharay (2005) and demonstrate the PPMC method using some most common and important statistics.

Observed sum scores

In many tests, ranking students is one of the most important objective. A model that does not predict the number of correct responses given by students well is very unlikely could rank students fairly. For the 5 dichotomous items in the example, there are 6 possible score categories. We observe a number of students within each score category. So the observed distribution of sum scores are $\boldsymbol{NC} = (NC_0, NC_1, \ldots, NC_5)$. Using the posterior samples produced by the Gibbs sampler, we can generate replicated datasets and compute the posterior predictive distribution of the sum scores. The result is in Figure 3

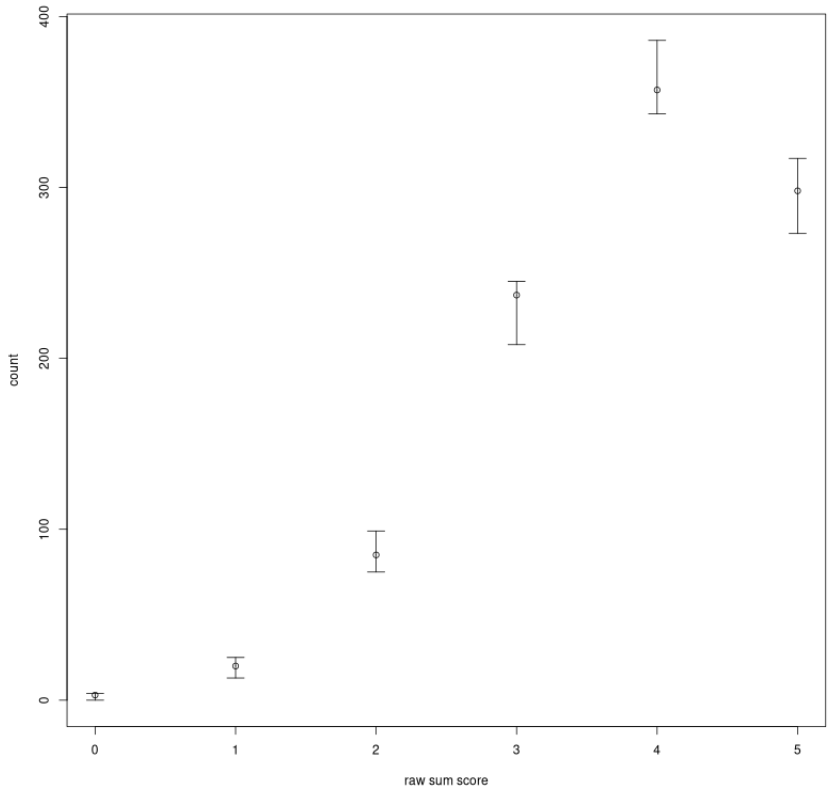

Figure 3

The observed sum scores and the predicted sum scores distribution.

The observed sum scores all fall within the 80% HPD intervals of the posterior predictive distributions. The model does seem to be able to explain the distribution of number of correct responses well.

Polychoric correlation coefficient

An important assumption of the 1PNO IRT model is that discrimination parameters across all items are equal. To check this assumption we use the polychoric correlation coefficient. A polychoric correlation coefficient describes the correlation between two observed ordinal variables - in our case, the total number of correct responses of a student and that student's item response to the j^{th} item. If the model is reasonable, then we would expect to see they polychoric correlations for all J items are similar or the variance of the polychoric correlations is close to 0. The results in Figure 4 shows the

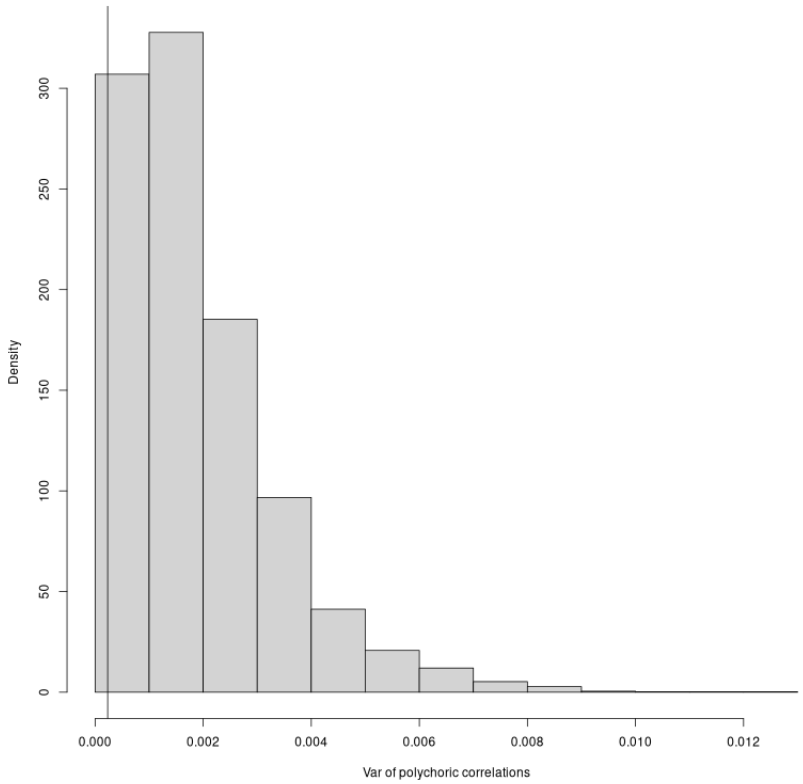

Figure 4

The observed and the predicted variance of polychoric correlations

variance of the observed polychoric correlations are very close to 0 and within a high density area of the posterior predictive distribution of the same statistic. Consequently, there is no evidence of this aspect of misfit of the model.

Odds ratio

The 1PNO IRT model assumes that item responses are independent of each other given latent abilities $\boldsymbol{\theta}$. This is also commonly referred to as the local independence (LI) assumption. If LI holds, the model should predict the odds ratio of any item pairs,

$$OR = \frac{n_{00}n_{11}}{n_{01}n_{10}}, \tag{6}$$

where $n_{kk'}$ denotes the number of students has a score k on the first item and a score k' on the second item. For 5 items, there are $\binom{5}{2} = 10$ pairs. For demonstration purpose, we plot

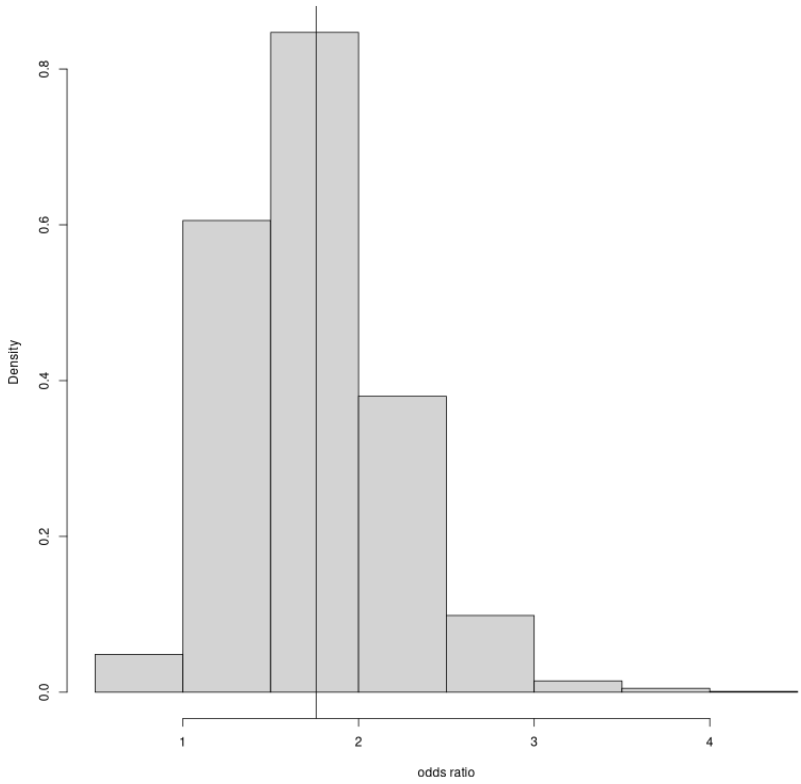

Figure 5

The observed and the predicted odds ratio

the results of one item pair in Figure 5. The observed OR is within a high density region of the posterior predictive distribution. The examined statistic does not show evidence of violation of the LI assumption for this pair of items.

Discussion

We introduced basic concepts of Bayesian psychometric modeling such as priors and hyperpriors, posterior computation, and posterior predictive model checking, all through a relatively simple IRT model example. For a more detailed introduction to these technical Bayesian topics, readers may refer to Gelman et al. (2013).

References

Bock, D. R., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items [ISBN: 978-3-642-85578-8]. *Psychometrika*, *35*(2), 179–197. https://doi.org/10.1007/BF02291262

Culpepper, S. A. (2015). Bayesian Estimation of the DINA Model With Gibbs Sampling [Publisher: SAGE PublicationsSage CA: Los Angeles, CA ISBN: 1076998615]. *Journal of Educational and Behavioral Statistics*, *40*(5), 454–476. https://doi.org/10.3102/1076998615595403

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2). https://doi.org/10.1007/BF02294839

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* [Publication Title: Book]. Chapman; Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images [ISBN: 0934613338]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Liu, X., & Johnson, M. S. (2019). Estimating CDMs Using MCMC. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 629–646). Springer. https://doi.org/10.1007/978-3-030-05584-4_31

Maris, G., Bechger, T., & San Martin, E. (2015). A Gibbs Sampler for the (Extended) Marginal Rasch Model. [Publisher: Springer]. *Psychometrika*, *80*(4), 859–79. https://doi.org/10.1007/s11336-015-9479-4

Neal, R. M. (1998). Probabilistic Inference Using Markov Chain Monte Carlo Methods [ISBN: 9780123736468]. *Technical Report*, *1*, 1–144. https://doi.org/10.1021/np100920q

Rizopoulos, D. (2015). Ltm : An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician [arXiv: 1306.3979v1 ISBN: 0090-5364]. *The Annals of Statistics*, *12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

Sinharay, S. (2005). Assessing Fit of Unidimensional Item Response Theory Models Using a Bayesian Approach [Publisher: Blackwell Publishing]. *Journal of Educational Measurement*, *42*(4), 375–394. https://doi.org/10.1111/j.1745-3984.2005.00021.x