ChinaRxiv

English-access research papers from ChinaXiv

A Comparison of the Accuracy of Variance-Component Estimation in Generalizability Theory Based on Three MCMC Sampling Methods[*]

Li Guangming

2026-06-17

ChinaRxiv

View the original and related papers at
https://chinarxiv.org/items/chinaxiv-202606.00117

Source: ChinaXiv. Machine translation. Verify with the original.

ChinaRxiv

Abstract

Generalizability theory is a modern psychometric theory concerning the reliability of behavioral measurement and has been widely applied in psychometric practice. Based on p × i and p × i × h designs, simulation and empirical studies were conducted to estimate variance components in generalizability theory using three MCMC sampling methods (the M-H algorithm, Gibbs sampling, and the HMC algorithm), and their differences were compared. The results showed that: (1) a small amount of missing data had almost no effect on the three MCMC sampling methods, and MCMC demonstrated strong robustness in estimating missing data; (2) overall, informative priors and empirical priors yielded higher estimation accuracy than noninformative priors; (3) only the HMC algorithm exhibited "cross-design" applicability, with high estimation accuracy under both designs, whereas the M-H algorithm and Gibbs sampling were limited by the estimation of variance components for the h facet and showed low estimation accuracy; (4) if the M-H algorithm and Gibbs sampling are to overcome the limitation of low accuracy in estimating variance components for the h facet, it is necessary to specify h $\geq$ 4. Future researchers are therefore recommended to use the HMC algorithm when applying MCMC methods.

Full Text

A Comparison of the Accuracy of Variance-Component Estimation in Generalizability Theory Based on Three MCMC Sampling Methods*

Li Guangming

(School of Psychology, Center for Psychological Application Research, South China Normal University, Guangzhou 510631)

Abstract Generalizability theory is a modern theory concerning the reliability of behavioral measurement and has been widely applied in psychometric practice. Based on the $p \times i$ and $p \times i \times h$ designs, this study used simulation research and empirical research, employing three MCMC sampling methods—the M-H algorithm, Gibbs sampling, and the HMC algorithm—to estimate variance components in generalizability theory and compare their differences. The results showed that: (1) a small amount of missing data had almost no effect on the three MCMC sampling methods, indicating that MCMC has strong robustness for estimation with missing data; (2) overall, informative priors and empirical priors yielded higher estimation accuracy than non-informative priors; (3) only the HMC algorithm demonstrated "cross-design applicability," achieving relatively high estimation accuracy under both designs, whereas the M-H algorithm and Gibbs sampling were limited by the estimation of the variance component for h, and their estimation accuracy was not high; (4) if the M-H algorithm and Gibbs sampling are to overcome the limitation of low accuracy in

estimating the variance component for h, it is necessary to stipulate $h \geq 4$; for future researchers using MCMC methods, the HMC algorithm is more highly recommended.

Keywords generalizability theory; variance-component estimation; MCMC methods; M-H algorithm; Gibbs sampling; HMC algorithm

1 Introduction

Generalizability Theory (GT) developed on the basis of classical test theory and is a theory concerning the reliability of behavioral measurement (Li Guangming & Qin Yue, 2022). GT can distinguish multiple sources of error in the measurement process and estimate the effects of errors from different sources on the object of measurement. It has been widely applied in psychometric practice, including cut-score setting, academic proficiency testing, performance assessment, personnel selection and assessment, and many other areas (Jackson et al., 2016; Li & Wu, 2023; Zhao Xuan et al., 2018; Zheng Xueqin et al., 2025).

Generalizability theory analysis includes two processes: the generalizability study (G study) and the decision study (D study) (Liao, 2023; Yang Zhiming, 2003). The main task of a G study is to identify as many sources of error as possible in the measurement process and to estimate the variance components (Variance Components, VC) of different error sources; the main task of a D study is, on the basis of the generalizability study, to control measurement error by adjusting the various relationships in the measurement process according to specific practical decision-making needs. It can thus be seen that variance

Date received: 2025-06-16

* General Program of the Natural Science Foundation of Guangdong Province (2021A1515012516), Guangdong Provincial Philosophy and Social Sciences Planning Discipline Co-construction Project (GD24XXL03), and the South China Normal University Peak Discipline Project for Strengthening Psychology.

Corresponding author: Li Guangming, E-mail: Lgm2004100@m.scnu.edu.cn

Estimation of variance components is a key step in GT analysis (Li Guangming et al., 2022).

In generalizability studies, the estimation of variance components is affected by many factors, including the estimation method, data distribution, research design, and sample size (Brennan et al., 2022; Li Guangming, 2016; Li, 2023). Methods such as Traditional, Bootstrap, Jackknife, MINQUE, REML (Restricted Maximum Likelihood Estimation), and MCMC (Markov Chain Monte Carlo) can all be used to estimate variance components (Brennan, 2001; Jackson et al., 2016; Lopilato et al., 2015). Previous studies have shown that applying MCMC methods to the estimation of variance components in GT has the following advantages (Chen et al., 2022; Enders et al., 2013; Jiang

& Skorupski, 2018; Lopilato et al., 2015): first, it can improve the accuracy of variance-component estimation and estimate confidence intervals for the generalizability coefficient and reliability index in D studies; it can also provide standard errors and confidence-interval estimates for covariance components in multivariate generalizability theory; second, compared with other estimation methods, the MCMC method does not estimate negative variance components; third, the MCMC method is flexible and, in theory, can better handle missing data and small-sample data.

However, there are multiple sampling methods within the MCMC approach, and they need to be compared. The Metropolis–Hastings (M-H) algorithm is the core sampling method in MCMC methods (Huang Changquan, 2017). Gibbs sampling is an M-H algorithm with an acceptance probability of 1, and it is simple and easy to understand (Liu Jinshan & Xia Qiang, 2016). Both the M-H algorithm and Gibbs sampling are local-updating algorithms and have unavoidable random-walk behavior; when estimating parameters in complex models, both algorithms may involve excessive computation and excessive time consumption, and especially when there are strong correlations among model parameters, their convergence is less than ideal (Hoffman & Gelman, 2014). Later, scholars proposed a new sampling method—the HMC (Hamiltonian Monte Carlo) algorithm—which can effectively avoid random-walk behavior in the sampling process and improve the efficiency of convergence (Neal, 2010). Studies have shown that, compared with the M-H algorithm and Gibbs sampling, the HMC algorithm has certain advantages (Nugroho & Morimoto, 2015). In addition to being a global-updating algorithm that can effectively avoid random-walk behavior, the HMC algorithm has a relatively high acceptance probability, can rapidly converge to the target posterior distribution, and can handle complex models with high-dimensional integration. Lopilato et al. (2015) were the first to use the HMC algorithm to estimate variance components in GT, and they used a two-facet crossed design to compare the estimation accuracy of the REML method and the MCMC method. However, they did not compare the estimation performance of the HMC algorithm with that of other MCMC sampling methods. At present, research on the HMC algorithm has mainly focused on statistics and on certain specific applied studies (Bourabee & Sanzserna, 2017; Chaudhuri et al., 2017; Monnahan et al., 2017; Zhang et al., 2017). Nevertheless, two shortcomings remain: first, few scholars have transferred the HMC algorithm to complex GT analyses with missing data; second, how do the three MCMC sampling methods perform when applied to GT analysis? To date, no scholars have explored this issue.

A one-facet crossed design refers to a test design with only one measurement facet, in which there is a crossed relationship between the object of measurement and the measurement facet; that is, every level of the object of measurement must be combined with every level of the measurement facet. If the object of measurement is p and the measurement facet is i, then a one-facet crossed design can be denoted by $p \times i$. A two-facet crossed design contains two measurement facets, and every level of each measurement facet must be paired one-to-one with every level of the object of measurement. For example, in a mathematics test for ninth-grade students, each student must

each student was required to complete 5 constructed-response items, and each student' s 5 items were rated by 3 mathematics teachers; that is, each teacher scored every student on each constructed-response item. If the object of measurement is p (students), and the two facets of measurement are i (items) and h (teachers), then the two-facet crossed design can be denoted as $p \times i \times h$.

This study aims to apply MCMC methods within the Bayesian theoretical framework to GT analysis, and to compare the performance differences of three sampling methods—the M-H algorithm, Gibbs sampling, and the HMC algorithm—in estimating variance components under different proportions of missing data. Specifically, based on two common generalized designs in GT (the $p \times i$ design and the $p \times i \times h$ design), this study uses simulation and empirical studies to examine the accuracy of the three MCMC sampling methods in estimating variance components, and further investigates their specific performance and applicable ranges, so as to provide theoretical guidance for corresponding practice.

2 Single-Facet Crossed Design ($p \times i$) Simulation Study

In the single-facet crossed design $p \times i$, the M-H algorithm, Gibbs sampling, and the HMC algorithm were used separately to estimate the variance components under each condition, and the performance of the three sampling methods in estimating variance components and their influencing factors were examined.

2.1 Research Design

Referring to previous settings for the single-facet crossed design $p \times i$ (Mao et al., 2005), the true values of the variance components $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$ were set to 4, 16, and 64, respectively. Normally distributed data were simulated, with the sample size fixed as follows: the number of levels of p was 100, and the number of levels of i was 20.

The independent variables were set as follows:

(1) Three MCMC sampling methods, namely the M-H algorithm, Gibbs sampling, and the HMC algorithm.

(2) Three types of prior information, namely informative-prior MCMC (MCMC with informative priors, MCMC inf), non-informative-prior MCMC (MCMC with non-informative priors, MCMC non), and empirical-prior MCMC (MCMC with empirical priors, MCMC emp).

According to the properties of conjugate prior distributions in Bayesian statistics (Huang Changquan, 2017), the target distribution of the prior information was set as the inverse-gamma distribution (IGamma). The settings of prior information followed previous studies (Lopilato et al., 2015; Mao et al., 2005). Table 1 presents the specific settings for the three types of prior information.

Table 1 Three Types of Prior Information for the Single-Facet Crossed Design $p \times i$

Prior information	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
MCMC inf	IGamma(2, 4)	IGamma(2, 16)	IGamma(2, 64)
MCMC non	IGamma(0.001, 0.001)	IGamma(0.001, 0.001)	IGamma(0.001, 0.001)
MCMC emp	IGamma(2, 5.95)	IGamma(2, 24.07)	IGamma(2, 95.97)

(3) Three proportions of completely random missingness, namely 0%, 5%, and 10%. In the practice of psychological statistics and measurement, data are often encountered

missingness. Prior research indicates that MCMC methods can better handle missing data (Enders et al., 2013). Therefore, with reference to the missingness proportions set by Lopilato et al. (2015) in their GT study, three missing-completely-at-random proportions were specified: 0%, 5%, and 10%.

In summary, the total number of simulation conditions was: 3 sampling methods (M-H, Gibbs, HMC) × 3 types of prior information (MCMC inf, MCMC non, MCMC emp) × 3 missingness proportions (0%, 5%, 10%) = 27. Under each condition, 1,000 simulations were conducted; that is, 1,000 sets of matrix data were generated. The estimated value of each variance component was taken as the mean of the 1,000 estimation results. Three variance components were to be estimated, namely $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$.

2.2 Comparison Criteria

The comparison criteria selected were estimation bias (Bias) and root mean square error (Root Mean Square Error, RMSE) (Medvedev & Siegert, 2022; Medvedeva et al., 2022). The calculation formulas are as follows:

$$Bias = \frac{\sum(\hat{\theta} - \theta)}{n} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum(\hat{\theta} - \theta)^2}{n}} \tag{2}$$

In formulas (1) and (2), $\hat{\theta}$ denotes the estimated value of the statistic in each simulation, that is, the estimated value of the variance component; θ denotes the true value of the statistic, that is, the true value of the variance component; and n denotes the number of simulations under each condition. The smaller the absolute value of Bias, the more accurate the estimation. The smaller the RMSE value, the more accurate the estimation.

2.3 Data Simulation and Estimation

Using Monte Carlo data simulation techniques, 1,000 sets of normally distributed matrix data were generated in R. The mathematical model on which this was based is as follows:

$$X_{pi} = \mu + \sigma_p z_p + \sigma_i z_i + \sigma_{pi} z_{pi} \tag{3}$$

In formula (3), the true values of the variance components $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$ were set to 4, 16, and 64, respectively, and μ was set to 0. The standard normal distributions z_p, z_i, and z_{pi} were simulated and multiplied by σ_p, σ_i, and σ_{pi}, respectively; the three products were then summed according to the formula to obtain the simulated data (Mao et al., 2005). On this basis, a complete data matrix with a sample size of 100×20 was generated. Then, missing-data matrices under missing completely at random were simulated according to the three different missingness proportions.

The M-H algorithm, Gibbs sampling, and the HMC algorithm were used to estimate the $p \times i$ design data, respectively, yielding the corresponding variance-component results. Specifically, the M-H algorithm was implemented by calling the BRugs package in R; Gibbs sampling was implemented by calling the R2jags package in R and linking it with JAGS; and the HMC algorithm was implemented by calling the Rstan package in R. For each MCMC algorithm, three Markov chains were set, each with 10,000 iterations. Because the random numbers generated at the very beginning of iteration are unstable, the first 5,000 iterations were discarded as burn-in results in this study.

2.4 Estimation Results

For the $p \times i$ design, the model was sampled using R software and ultimately reached convergence. By observing the historical iteration plots, setting the burn-in number to 5000 was sufficient for the chain to reach a stationary distribution. Therefore, from iterations 5001 to 10000, the values of each Markov chain were taken as estimates, and the average of the three Markov chains was used as the estimate for each statistic. The estimation results for the variance components of the $p \times i$ design under the three missing-data proportions (missing 0%/5%/10%) are shown in Table 2.

Table 2. Estimation results for variance components under the $p \times i$ design (missing 0%/5%/10%)

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
		True value	4	16	64

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
MCMC inf	M-H	VC	3.797/3.815/3.743	15.664/15.727/15.868	64.059/64.121/64.167
MCMC inf	M-H	Bias	-0.203/-0.185/-0.257	-0.336/-0.273/-0.132	0.059/0.121/0.167
MCMC inf	M-H	RMSE	0.938/0.980/1.029	4.679/4.966/4.937	2.072/2.167/2.270
MCMC inf	Gibbs	VC	3.844/3.754/3.801	15.873/15.971/15.741	64.110/64.197/64.033
MCMC inf	Gibbs	Bias	-0.156/-0.246/-0.199	-0.127/-0.029/-0.259	0.110/0.197/0.033
MCMC inf	Gibbs	RMSE	1.025/0.989/1.059	4.929/5.082/4.725	2.106/2.123/2.264
MCMC inf	HMC	VC	3.625/3.694/3.683	15.156/15.112/15.199	64.124/64.095/64.173
MCMC inf	HMC	Bias	-0.375/-0.306/-0.317	-0.844/-0.888/-0.801	0.124/0.095/0.173
MCMC inf	HMC	RMSE	0.991/1.021/1.083	4.735/4.741/4.721	2.160/2.196/2.250
MCMC non	M-H	VC	4.015/4.014/4.013	17.664/17.683/17.835	63.966/64.156/64.175
MCMC non	M-H	Bias	0.015/0.014/0.013	1.664/1.683/1.835	-0.034/0.156/0.175
MCMC non	M-H	RMSE	1.068//1.068/1.068	6.023/6.219/6.321	2.047/2.162/2.189
MCMC non	Gibbs	VC	4.065/4.048/3.957	17.787/17.678/18.031	64.063/64.138/64.114
MCMC non	Gibbs	Bias	0.065/0.048/-0.043	1.787/1.678/2.031	0.063/0.138/0.114
MCMC non	Gibbs	RMSE	1.086/1.079/1.138	6.261/6.086/6.501	2.037/2.175/2.272
MCMC non	HMC	VC	3.897/3.813/3.763	16.390/16.909/16.758	64.111/64.074/64.190
MCMC non	HMC	Bias	-0.103/-0.187/-0.237	0.390/0.909/0.758	0.111/0.074/0.190
MCMC non	HMC	RMSE	1.132/1.157/1.215	5.585/5.914/5.725	2.470/2.157/2.214
MCMC emp	M-H	VC	3.980/3.931/3.963	16.639/16.760/16.814	64.048/64.155/64.022
MCMC emp	M-H	Bias	-0.020/-0.069/-0.037	0.639/0.760/0.814	0.048/0.155/0.022
MCMC emp	M-H	RMSE	0.964/0.908/0.943	4.747/4.851/4.921	2.016/2.162/2.185

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
MCMC emp	Gibbs	VC	3.990/3.900/3.898	16.449/16.836/16.61[illegible]	64.096/64.056/64.195
MCMC emp	Gibbs	Bias	-0.010/-0.100/-0.102	0.449/0.836/0.6[illegible]	0.096/0.056/0.195
MCMC emp	Gibbs	RMSE	0.951/0.943/0.965	4.512/4.917/4.82[illegible]	2.264/2.177/2.236
MCMC emp	HMC	VC	3.836/3.818/3.853	15.053/15.814/16.032	64.111/64.111/64.182
MCMC emp	HMC	Bias	-0.164/-0.182/-0.147	0.053/-0.186/0.032	0.111/0.111/0.182
MCMC emp	HMC	RMSE	0.951/0.977/0.997	4.843/4.587/4.520	2.034/2.114/2.152

According to the information in Table 2, the estimation results of the variance components, Bias, and RMSE for different missing-data proportions under the three prior-information conditions can be obtained for the M-H algorithm, Gibbs sampling, and the HMC algorithm. Considering that when the missing-data proportions are 0% and 5%, their results are consistent with those for a missing-data proportion of 10%

The results were basically the same and are not listed here; taking only the missing proportion of 10% as an example, the results are shown in Figure 1.

Figure 1. Bias and RMSE plots for the variance components of the $p \times i$ design (missing proportion 10%)

(a) Missing 10%, MCMC inf

(b) Missing 10%, MCMC non

(c) Missing 10%, MCMC emp

(d) Missing 10%, MCMC inf

(e) Missing 10%, MCMC non

(f) Missing 10%, MCMC emp

From the Bias results, under the informative-prior condition (Figure 1a), the estimation-bias ranges of the M-H algorithm, Gibbs sampling, and the HMC algorithm are $(-1, 0.4)$; under the empirical-prior condition (Figure 1c), the estimation-bias ranges of the three are $(-0.2, 1)$; and under the noninformative-prior condition (Figure 1b), the estimation-bias ranges of the three are $(-0.5, 2)$.

Under the three prior-information conditions, the estimation biases of the M-H algorithm, Gibbs sampling, and the HMC algorithm for $\sigma^2(p)$ are all relatively close, but their estimation-bias performance for $\sigma^2(i)$ and $\sigma^2(pi)$ is not the same. Under the informative-prior condition, the estimation bias of the M-H algorithm for $\sigma^2(i)$ is close to 0, whereas the estimation bias of Gibbs sampling for $\sigma^2(pi)$ is closer to 0; under the empirical-prior condition, the estimation bias of the HMC algorithm for $\sigma^2(i)$ is close to 0, whereas the estimation bias of the M-H algorithm for $\sigma^2(pi)$ is close to 0; under the noninformative-prior condition, the estimation biases of the three sampling methods for $\sigma^2(pi)$ are relatively close, while the estimation bias of the HMC algorithm for $\sigma^2(i)$ is slightly smaller.

From the RMSE results, when the missing proportion is 10%, under the three prior-information conditions, the RMSEs of the M-H algorithm, Gibbs sampling, and the HMC algorithm for estimating the three variance components are all relatively close. Among them, the RMSE values for $\sigma^2(p)$ are all close to 1, the RMSE values for $\sigma^2(pi)$ are all close to 2, and the RMSE values for $\sigma^2(i)$ are close to 5 under the informative-prior and empirical-prior conditions, and closer to 6 under the noninformative-prior condition.

In summary, in the one-facet crossed design $p \times i$, the variance-component estimation accuracy of the M-H algorithm, Gibbs sampling, and the HMC algorithm is relatively consistent. Under different missing proportions, the performance of variance-component estimation differs very little. Under the three prior-information conditions, the differences in the accuracy of variance-component estimation are small, and the estimation accuracy of the variance components is relatively higher under the informative-prior and empirical-prior conditions.

3 Simulation Study of the Two-Facet Crossed Design ($p \times i \times h$)

In the two-facet crossed design $p \times i \times h$, the M-H algorithm, Gibbs sampling, and the HMC algorithm were used respectively to estimate the variance components under each condition, in order to examine the performance of the three sampling methods in estimating variance components and the factors influencing their performance.

3.1 Research Design

Referring to previous settings for the two-facet crossed design $p \times i \times h$ (Tong & Brennan, 2007; Wang Xingjun et al., 2016), the true values of the variance components $\sigma^2(p)$, $\sigma^2(i)$, $\sigma^2(h)$, $\sigma^2(pi)$, $\sigma^2(ph)$, $\sigma^2(ih)$, and $\sigma^2(pih)$ were set to 16, 4, 1, 64, 2, 3, and 144, respectively, and normally distributed data were simulated. The sample size was fixed as follows: the number of levels of p was 100, the number of levels of i was 20, and the number of levels of h was 2. According to previous research (Lopilato et al., 2015), the setting values for the informative priors were the true values of the variance components; the setting values for the noninformative priors were IGamma(0.001, 0.001); and the setting values for the empirical priors were 1.5 times the estimates of the variance components in the two-facet crossed design obtained by the REML method, as shown in Table 3.

Table 3 Three Types of Prior Information for the Two-Facet Crossed Design $p \times i \times h$

Variance component	MCMC inf	MCMC non	MCMC emp
$\sigma^2(p)$	IGamma(2, 16)	IGamma(0.001, 0.001)	IGamma(2, 23.76)
$\sigma^2(i)$	IGamma(2, 4)	IGamma(0.001, 0.001)	IGamma(2, 6.03)
$\sigma^2(h)$	IGamma(2, 1)	IGamma(0.001, 0.001)	IGamma(2, 1.53)
$\sigma^2(pi)$	IGamma(2, 64)	IGamma(0.001, 0.001)	IGamma(2, 95.93)
$\sigma^2(ph)$	IGamma(2, 2)	IGamma(0.001, 0.001)	IGamma(2, 3.02)
$\sigma^2(ih)$	IGamma(2, 3)	IGamma(0.001, 0.001)	IGamma(2, 4.52)
$\sigma^2(pih)$	IGamma(2, 144)	IGamma(0.001, 0.001)	IGamma(2, 216.14)

As in the one-facet crossed design $p \times i$, the missing-data proportions for the two-facet crossed design $p \times i \times h$ were set to 0%, 5%, and 10%, yielding a total of 27 simulation conditions. Under each condition, 1000 simulations were conducted, and the mean of the 1000 estimation results was taken as the estimate of the variance component. In total, 7 variance components needed to be estimated, namely $\sigma^2(p)$, $\sigma^2(i)$, $\sigma^2(h)$, $\sigma^2(pi)$, $\sigma^2(ph)$, $\sigma^2(ih)$, and $\sigma^2(pih)$.

3.2 Comparison Criteria

The same as formulas (1) and (2).

3.3 Data Simulation and Estimation

In R software, 1000 batches of normally distributed matrix data were generated. The mathematical model on which they were based is as follows:

$$X_{pi} = \mu + \sigma_p z_p + \sigma_i z_i + \sigma_h z_h + \sigma_{pi} z_{pi} + \sigma_{ph} z_{ph} + \sigma_{ih} z_{ih} + \sigma_{pih} z_{pih} \quad (4)$$

From formula (4), it can be seen that seven normally distributed values need to be simulated, namely z_p, z_i, z_h, z_{pi}, z_{ph}, z_{ih}, and z_{pih}; these seven values are respectively

After multiplying by the corresponding true values of the variance components, $\sigma^2(p)$, $\sigma^2(i)$, $\sigma^2(h)$, $\sigma^2(pi)$, $\sigma^2(ph)$, $\sigma^2(ih)$, and $\sigma^2(pih)$, namely (16, 4, 1, 64, 2, 3, 144), and summing them, the simulated random numbers are obtained, generating a complete data matrix with sample size $100 \times 20 \times 2$. Then, according to three different missingness proportions, missing-data matrices under completely random missingness are simulated. Subsequently, the M-H algorithm, Gibbs sampling, and the HMC algorithm are used separately to estimate the $p \times i \times h$ design data, yielding the corresponding estimates of variance components. Among them, the M-H algorithm is implemented by calling the BRugs package in R; Gibbs sampling is implemented by calling R2OpenBUGS in R and linking it with BUGS software; and the HMC algorithm is implemented by calling the Rstan package in R. For each sampling method, three Markov chains are set up; each chain is iterated 10,000 times, and the first 5,000 iterations are discarded.

3.4 Estimation Results

3.4.1 Estimation Results under Different Missingness Proportions As with the one-facet crossed $p \times i$ design, sampling is performed in R for the generalized model of the $p \times i \times h$ design, and convergence is ultimately achieved. By observing the historical iteration plots, setting the burn-in number to 5,000 is sufficient to reach a stationary distribution. Therefore, from iterations 5,001 to 10,000, the values of each Markov chain are taken as estimates, and the average over the three Markov chains is used as the estimate of each statistic. The estimation results for the variance components of the $p \times i \times h$ design (missingness 0%, 5%, and 10%) are shown in Tables 4 to 6.

Table 4 Estimation results of variance components for the $p \times i \times h$ design (missingness 0%)

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
		True value	16	4	1	64	2	3	144
MCMC inf	M-H	VC	15.502	3.635	0.998	63.292	1.641	2.980	145.044
MCMC inf	M-H	Bias	-0.498	-0.365	-0.002	-0.708	-0.359	-0.020	1.044
MCMC inf	M-H	RMSE	3.214	1.566	0.419	5.126	0.804	1.142	4.801
MCMC inf	Gibbs	VC	15.598	3.785	0.969	63.422	1.733	3.032	144.715

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC inf	Gibbs	Bias	-0.402	-0.215	-0.031	-0.578	-0.267	0.032	0.715
MCMC inf	Gibbs	RMSE	3.420	1.599	0.377	4.992	0.777	1.159	4.752
MCMC inf	HMC	VC	15.391	3.357	0.718	63.386	1.487	2.784	144.717
MCMC inf	HMC	Bias	-0.609	-0.643	-0.282	-0.614	-0.513	-0.216	0.717
MCMC inf	HMC	RMSE	3.219	1.644	0.446	5.160	0.852	1.177	4.600
MCMC non	M-H	VC	16.648	3.811	56.566	63.578	1.613	3.778	144.613
MCMC non	M-H	Bias	0.648	-0.189	55.566	-0.422	-0.387	0.778	0.613
MCMC non	M-H	RMSE	3.648	2.757	64.149	5.043	1.355	2.027	4.594
MCMC non	Gibbs	VC	16.506	3.656	159.590	63.320	1.464	3.867	144.886
MCMC non	Gibbs	Bias	0.506	-0.344	158.590	-0.680	-0.536	0.867	0.886
MCMC non	Gibbs	RMSE	3.627	2.733	593.552	5.001	1.338	2.164	4.773
MCMC non	HMC	VC	16.426	1.923	2.083	62.768	0.601	4.750	146.259
MCMC non	HMC	Bias	0.426	-2.077	1.083	-1.232	-1.399	1.750	2.259

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
[[unclear: continued from previous page]]	[[unclear: continued from previous page]]	RMSE	3.489	3.251	5.550	5.733	1.767	3.134	5.772
MCMC-emp	M-H	VC	15.865	4.212	1.412	63.631	1.977	3.124	144.620
MCMC-emp	M-H	Bias	-0.135	0.212	0.412	-0.369	-0.023	0.124	0.620
MCMC-emp	M-H	RMSE	3.136	1.494	0.631	4.828	0.745	1.042	4.685
MCMC-emp	Gibbs	VC	15.919	4.115	1.376	63.497	2.025	3.116	144.574
MCMC-emp	Gibbs	Bias	-0.081	0.115	0.376	-0.503	0.025	0.116	0.574
MCMC-emp	Gibbs	RMSE	3.183	1.511	0.583	4.850	0.716	1.031	4.547
MCMC-emp	HMC	VC	15.663	3.817	1.012	63.511	1.823	2.946	144.696
MCMC-emp	HMC	Bias	-0.337	-0.183	0.012	-0.489	-0.177	-0.054	0.696
MCMC-emp	HMC	RMSE	3.207	1.383	0.320	5.098	0.688	1.002	4.683

Table 5 Estimation results of variance components for the $p \times i \times h$ design (5% missing)

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
		True value	16	4	1	64	2	3	144

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC-inf	M-H	VC	15.722	3.767	0.949	63.682	1.664	2.871	144.847
MCMC-inf	M-H	Bias	-0.278	-0.233	-0.051	-0.318	-0.336	-0.129	0.847
MCMC-inf	M-H	RMSE	3.210	1.629	0.425	5.351	0.770	1.126	5.046
MCMC-inf	Gibbs	VC	15.855	3.642	0.988	63.427	1.756	2.869	144.797
MCMC-inf	Gibbs	Bias	-0.145	-0.358	-0.012	-0.573	-0.244	-0.131	0.797
MCMC-inf	Gibbs	RMSE	3.427	1.571	0.462	5.270	0.784	1.123	4.853
MCMC-inf	HMC	VC	15.285	3.389	0.729	63.232	1.519	2.791	144.992
MCMC-inf	HMC	Bias	-0.715	-0.611	-0.271	-0.768	-0.481	-0.209	0.992
MCMC-inf	HMC	RMSE	3.423	1.582	0.445	5.159	0.819	1.130	4.942
MCMC-non	M-H	VC	16.639	3.554	91.233	63.441	1.350	3.815	145.196
MCMC-non	M-H	Bias	0.639	-0.446	90.233	-0.559	-0.650	0.815	1.196
MCMC-non	M-H	RMSE	3.558	2.630	213.273	5.287	1.353	2.104	5.082
MCMC-non	Gibbs	VC	16.495	3.392	62.578	63.518	1.453	3.934	145.250
MCMC-non	Gibbs	Bias	0.495	-0.608	61.578	-0.482	-0.547	0.934	1.250
MCMC-non	Gibbs	RMSE	3.601	2.710	265.197	5.258	1.427	2.130	5.108
MCMC-non	HMC	VC	16.476	2.099	2.016	62.650	0.614	4.707	146.432
MCMC-non	HMC	Bias	0.476	-1.901	1.016	-1.350	-1.386	1.707	2.432
MCMC-non	HMC	RMSE	3.672	3.243	6.316	6.733	1.782	3.204	6.582
MCMC-emp	M-H	VC	15.777	4.215	1.312	63.886	1.996	3.033	144.508
MCMC-emp	M-H	Bias	-0.223	0.215	0.312	-0.114	-0.004	0.033	0.508

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC-emp	M-H	RMSE	3.299	1.557	0.531	5.277	0.743	1.023	5.091
MCMC-emp	Gibbs	VC	15.901	4.127	1.337	63.873	2.059	3.143	144.427
MCMC-emp	Gibbs	Bias	-0.099	0.127	0.337	-0.127	0.059	0.143	0.427
MCMC-emp	Gibbs	RMSE	3.183	1.418	0.517	5.105	0.754	1.086	4.905
	HMC	VC	15.648	3.811	1.010	63.684	1.825	2.949	144.390
	HMC	Bias	-0.352	-0.189	0.010	-0.316	-0.175	-0.051	0.390
	HMC	RMSE	3.245	1.423	0.328	5.403	0.718	1.051	4.972

Table 6 Estimation results for variance components under the $p \times i \times h$ design (10% missing)

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
		True value	16	4	1	64	2	3	144
MCMC-inf	M-H	VC	15.883	3.801	0.985	63.209	1.686	2.899	144.971
MCMC-inf	M-H	Bias	-0.117	-0.199	-0.015	-0.791	-0.314	-0.101	0.971
MCMC-inf	M-H	RMSE	3.421	1.667	0.458	5.427	0.808	1.131	5.252
MCMC-inf	Gibbs	VC	15.551	3.792	0.983	63.123	1.813	2.860	144.900
MCMC-inf	Gibbs	Bias	-0.449	-0.208	-0.017	-0.877	-0.187	-0.140	0.900
MCMC-inf	Gibbs	RMSE	3.281	1.643	0.453	5.393	0.860	1.142	5.148

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC inf	HMC	VC	15.428	3.297	0.719	63.256	1.478	2.691	145.132
MCMC inf	HMC	Bias	-0.572	-0.703	-0.281	-0.744	-0.522	-0.309	1.132
MCMC inf	HMC	RMSE	3.383	1.619	0.423	5.488	0.822	1.109	5.286
MCMC non	M-H	VC	16.583	3.509	49.424	63.045	1.304	3.955	145.767
MCMC non	M-H	Bias	0.583	-0.491	48.424	-0.955	-0.696	0.955	1.767
MCMC non	M-H	RMSE	3.632	2.705	94.872	5.647	1.425	2.297	5.450
MCMC non	Gibbs	VC	16.515	3.563	61.315	63.673	1.405	3.855	145.220
MCMC non	Gibbs	Bias	0.515	-0.437	60.315	-0.327	-0.595	0.855	1.220
MCMC non	Gibbs	RMSE	3.465	2.769	199.282	5.415	1.428	2.219	5.323
MCMC non	HMC	VC	16.520	2.113	2.070	62.654	0.568	4.713	146.663
MCMC non	HMC	Bias	0.520	-1.887	1.070	-1.346	-1.432	1.713	2.663
MCMC non	HMC	RMSE	3.698	3.280	5.760	6.691	1.786	3.321	6.786
MCMC emp	M-H	VC	15.830	4.230	1.330	63.785	2.020	3.053	144.306
MCMC emp	M-H	Bias	-0.170	0.230	0.330	-0.215	0.020	0.053	0.306
MCMC emp	M-H	RMSE	3.278	1.480	0.537	5.390	0.771	1.077	5.059
MCMC emp	Gibbs	VC	15.798	4.128	1.355	63.533	2.061	3.075	144.474
MCMC emp	Gibbs	Bias	-0.202	0.128	0.355	-0.467	0.061	0.075	0.474
MCMC emp	Gibbs	RMSE	3.368	1.502	0.553	5.525	0.756	1.086	5.203
MCMC emp	HMC	VC	15.572	3.815	0.995	63.379	1.831	2.908	144.690
MCMC emp	HMC	Bias	-0.428	-0.185	-0.005	-0.621	-0.169	-0.092	0.690

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC emp	HMC	RMSE	3.292	1.444	0.283	5.616	0.677	1.040	4.816

According to the information in Tables 4, 5, and 6, the estimation results for the variance components, Bias, and RMSE of the M-H algorithm, Gibbs sampling, and the HMC algorithm under the three prior-information conditions and different missingness proportions (0%, 5%, and 10%) can be obtained. As with the one-facet crossed $p \times i$ design, taking only the 10% missingness proportion as an example, the results are shown in Figure 2.

[Figure: Six line charts with legends "M-H," "Gibbs," and "HMC." The top row shows Bias and the bottom row shows RMSE for variance components p, i, h, pi, ph, ih, and pih.]

a. Missing 10%, MCMC inf

b. Missing 10%, MCMC non

c. Missing 10%, MCMC emp

d. Missing 10%, MCMC inf

e. Missing 10%, MCMC non

f. Missing 10%, MCMC emp

Figure 2 Bias and RMSE plots of variance components for the $p \times i \times h$ design (missing proportion 10%)

From the Bias results, under the informative-prior condition (panel a in Figure 2), the range of estimation bias for the M-H algorithm, Gibbs sampling, and the HMC algorithm is $(-1, 1.2)$. Under the empirical-prior condition (panel c in Figure 2), the range of estimation bias for the three methods is $(-0.6, 0.8)$. Under the noninformative-prior condition (panel b in Figure 2), however, the estimation bias of the M-H algorithm and Gibbs sampling for $\sigma^2(h)$ is very large, at 48.424 and 60.315, respectively. Excluding these two Bias values, the range of estimation bias for the three sampling methods is $(-2, 3)$. Under the informative prior, the estimation bias of the M-H algorithm for $\sigma^2(p)$ is closest to 0; the estimation biases of the M-H algorithm and Gibbs sampling for $\sigma^2(i)$, $\sigma^2(h)$, and $\sigma^2(ph)$ are slightly smaller than those of the HMC algorithm (but quite close), and the three methods have similar estimation biases for the remaining three variance components. Under the empirical prior, the estimation bias of the HMC algorithm for $\sigma^2(h)$ is closest to 0; the estimation biases of the M-H algorithm and Gibbs sampling for $\sigma^2(p)$, $\sigma^2(pi)$, $\sigma^2(ph)$, and $\sigma^2(pih)$ are slightly smaller than those of the HMC algorithm (but quite close), and the three methods have similar estimation biases for $\sigma^2(i)$ and $\sigma^2(ih)$.

From the RMSE results, under the informative-prior and empirical-prior conditions, the RMSEs of the variance-component estimates obtained by the M-H algorithm, Gibbs sampling, and the HMC algorithm are all basically similar. The three sampling algorithms have smaller RMSEs when estimating $\sigma^2(h)$, $\sigma^2(ph)$, and $\sigma^2(ih)$, and larger RMSEs when estimating $\sigma^2(p)$, $\sigma^2(pi)$, and $\sigma^2(pih)$. Under the noninformative prior, the RMSEs of the M-H algorithm and Gibbs sampling for estimating $\sigma^2(h)$ are very large, at 94.872 and 199.282, respectively.

In summary, except under the noninformative-prior condition, where the estimation bias of the M-H algorithm and Gibbs sampling for $\sigma^2(h)$ is excessively large, under the other conditions the estimates of the variance components obtained by the three sampling methods fluctuate around the true values. This indicates that, under the noninformative prior, only the HMC algorithm can accurately estimate the variance component $\sigma^2(h)$. Under the other conditions, the variance-component estimates of the three sampling methods are relatively close, although their specific performance differs.

3.4.2 Estimation results for different numbers of levels of facet h

Under the noninformative-prior condition, the estimation bias of the M-H algorithm and Gibbs sampling for $\sigma^2(h)$ is too large. Therefore, increasing the number of levels of facet h

number, in order to examine whether the M-H algorithm and Gibbs sampling can estimate $\sigma^2(h)$ more accurately. Under the non-informative prior, when the number of levels of the h facet is 3 to 6, the estimation results for $\sigma^2(h)$ obtained by the M-H algorithm and Gibbs sampling are shown in Table 7 and Table 8.

Table 7 Results for estimating $\sigma^2(h)$ with the M-H algorithm

Missing proportion	Estimate	$h = 3$	$h = 4$	$h = 5$	$h = 6$
Missing 0%	VC	3.782	2.200	1.965	1.505
Missing 0%	Bias	2.782	1.200	0.965	0.505
Missing 0%	RMSE	6.219	2.598	1.987	1.343
Missing 5%	VC	8.153	2.916	1.789	1.466
Missing 5%	Bias	7.153	1.916	0.789	0.466
Missing 5%	RMSE	12.026	3.397	1.925	1.268
Missing 10%	VC	6.809	2.817	1.782	1.533
Missing 10%	Bias	5.809	1.817	0.782	0.533
Missing 10%	RMSE	9.538	3.310	1.817	1.400

Table 8 Results for estimating $\sigma^2(h)$ with Gibbs sampling

Missing proportion	Estimate	$h=3$	$h=4$	$h=5$	$h=6$
Missing 0%	VC	8.136	2.965	1.853	1.637
Missing 0%	Bias	7.136	1.965	0.853	0.637
Missing 0%	RMSE	11.989	3.821	1.952	1.512
Missing 5%	VC	5.676	2.769	1.868	1.666
Missing 5%	Bias	4.676	1.769	0.868	0.666
Missing 5%	RMSE	7.781	3.332	1.888	1.543
Missing 10%	VC	6.195	2.526	1.916	1.533
Missing 10%	Bias	5.195	1.526	0.916	0.533
Missing 10%	RMSE	9.044	2.940	1.958	1.405

As can be seen from Table 7 and Table 8, as the number of levels of the h facet increases, the estimates of $\sigma^2(h)$ obtained by the M-H algorithm and Gibbs sampling become increasingly close to the true value of 1, and both the Bias and RMSE values gradually decrease. When the number of levels of the h facet is 3, the estimates of $\sigma^2(h)$ obtained by the M-H algorithm and Gibbs sampling remain relatively large. However, when the number of levels of the h facet increases to 4, the estimation bias of $\sigma^2(h)$ under the M-H algorithm and Gibbs sampling is smaller; when it further increases to 5 and 6, the estimation bias does not increase. When the number of levels of the h facet is 4–6, the range of Bias values for the variance component is relatively stable and is close to the ranges under the informative prior and the empirical prior. When the number of levels of the h facet increases to 4 or more, the estimation bias of $\sigma^2(h)$ under the M-H algorithm and Gibbs sampling is smaller and is relatively close to the results under the informative prior and the empirical prior. Therefore, if future researchers continue to use the M-H algorithm and Gibbs sampling to estimate variance components in two-facet crossed designs, the number of levels of the h facet in the study should not be less than 4, so as not to affect the accuracy of the estimation results.

4 Empirical Study

4.1 Empirical Data for the $p \times i$ Design

Students were measured using the Self-Rating Depression Scale (SDS) developed by Zung (1965). The SDS contains 20 items reflecting subjective feelings of depression, among which 10 items are positively scored and the other 10 are reverse-scored. Items include statements such as "I feel down-hearted and blue" and "Morning is when I feel the best." A four-point scale is used according to the frequency with which symptoms occur: 1 indicates "none or a little of the time," 2 indicates "some of the time," 3 indicates "a good part of the time," and 4 indicates "most or all of the time."

The data came from a random sample of 1,545 students in a city in southern China. They were followed up at intervals of 6 months, and three paper-and-pencil tests were administered in total. Attrition occurred in all three tests, possibly for the following reasons: (1) absence on the day of testing; (2) incorrect or missing responses to items; and (3) transfer to another school or suspension of schooling. After excluding participants who did not complete all three tests, 1,429 participants were retained, including 636 male adolescents and 793 female adolescents, with a mean age of 14.88 years (standard deviation = 1.80). The scale showed good internal consistency in this study (T1 Cronbach' s $\alpha = 0.902$; T2 Cronbach' s $\alpha = 0.906$; T3 Cronbach' s $\alpha = 0.922$).

Among the random sample of 1,429 adolescents, 1,288 participants had complete data and 141 participants had missing data, accounting for 9.9%. This design is a $p \times i$ design, where p denotes adolescents (1,429 in total), and i denotes the items of the depression scale (20 items in total). The item mean scores from the three measurements were taken as the item scores, namely the item mean scores at T1, T2, and T3. For the data of the 1,429 participants, 5,000 repeated estimations were first conducted, and the mean of the estimated variance components was regarded as the "parameter" (true value) (Tong & Brennan, 2007). Because no prior information on these data was available from previous studies, variance components were estimated using an MCMC method with noninformative priors. The results are shown in Table 9.

Table 9 Estimation Results of Variance Components for the $p \times i$ Design (9.9% Missing)

Prior information	Sampling method	Estimated value	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
	M-H	VC	4.005	17.800	65.102
	M-H	Bias	0.005	1.800	1.102
	M-H	RMSE	1.091	6.306	2.189
MCMC non	Gibbs	VC	3.999	17.089	64.199
MCMC non	Gibbs	Bias	-0.001	1.089	0.199
MCMC non	Gibbs	RMSE	1.189	2.501	2.236
MCMC non	HMC	VC	3.999	16.704	64.177
MCMC non	HMC	Bias	-0.001	0.704	0.177
MCMC non	HMC	RMSE	1.255	2.788	2.202

It can be seen from Table 9 that the results of the three sampling methods are all relatively good, which is consistent with the conclusions from the simulation results. Comparatively speaking, HMC is overall superior to the M-H and Gibbs methods.

4.2 Empirical Data for the $p \times i \times h$ Design

Using the SDS scale, repeated measurements were administered to 1,429 adolescents at a one-year interval. This design is a $p \times i \times h$ design, where p denotes adolescents (1,429 in total), i denotes the items of the depression scale (20 items in total), and h denotes repeated measurements at two time points (T1 and T3),

The missing rate was approximately 9.9%. As in the single-facet crossed $p \times i$ design, the data from 1,429 examinees were first estimated repeatedly 5,000 times, and the means of the estimated variance components were regarded as "parameters" (true values) (Tong & Brennan, 2007). Moreover, because previous researchers had no prior information about these data, variance components were estimated using MCMC methods with non-informative priors. The results are shown in Table 10.

Table 10 Estimation results for variance components in the $p \times i \times h$ design (missing 9.9%)

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC non-informative	M-H	VC	16.983	3.559	49.404	63.245	1.388	3.988	145.700
MCMC non-informative	M-H	Bias	0.983	-0.441	48.404	-0.755	-0.612	0.988	1.700
MCMC non-informative	M-H	RMSE	3.639	2.745	92.002	5.007	1.400	2.207	5.459
MCMC non-informative	Gibbs	VC	16.515	3.563	61.309	63.673	1.405	3.889	145.288
MCMC non-informative	Gibbs	Bias	0.515	-0.437	60.309	-0.327	-0.595	0.889	1.288
MCMC non-informative	Gibbs	RMSE	3.465	2.769	190.202	5.415	1.428	2.208	5.309

Prior information	Sampling method	Estimate	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(h)$	$\sigma^2(pi)$	$\sigma^2(ph)$	$\sigma^2(ih)$	$\sigma^2(pih)$
MCMC non-informative	HMC	VC	16.529	2.119	2.071	62.600	0.587	3.753	146.612
MCMC non-informative	HMC	Bias	0.529	-1.881	1.071	-1.400	-1.413	0.753	2.612
MCMC non-informative	HMC	RMSE	3.690	3.289	5.761	6.608	1.789	3.308	6.7822

As can be seen from Table 10, the larger biases are mainly reflected in $\sigma^2(h)$ for M-H and Gibbs, whereas $\sigma^2(h)$ under the HMC method remains relatively small. This is consistent with the conclusion from the simulation results.

5 General Discussion

5.1 Comparison of the Three MCMC Sampling Methods

In the simulation study of the single-facet crossed $p \times i$ design, the M-H algorithm, Gibbs sampling, and the HMC algorithm were all able to estimate variance components relatively accurately under different conditions. In the simulation study of the two-facet crossed $p \times i \times h$ design, under informative-prior and empirical-prior conditions, the estimation accuracy of the M-H algorithm, Gibbs sampling, and the HMC algorithm was comparable; under the non-informative-prior condition, however, the HMC algorithm yielded higher accuracy in estimating variance components. When the M-H algorithm and the Gibbs algorithm estimated the variance component $\sigma^2(h)$ of the two-facet crossed design under the non-informative-prior condition, both produced excessively large estimation bias. Whether in the single-facet crossed design or in the two-facet crossed design, the estimation results for variance components obtained by the M-H algorithm and Gibbs sampling were relatively consistent. Indeed, under the non-informative-prior condition in the two-facet crossed design, both showed a highly consistent bias in estimating $\sigma^2(h)$. The consistency displayed by the M-H algorithm and Gibbs sampling in estimating variance components is closely related to the principles of, and connections between, the two sampling methods.

In the empirical study, whether for the empirical data of the $p \times i$ design or for the empirical data of the $p \times i \times h$ design, the results were almost entirely consistent with the conclusions of the simulation study. In the empirical data for the $p \times i$ design, the results of the M-H method, Gibbs method, and HMC method were all relatively good; however, in the empirical data for the $p \times i \times h$ design, when the M-H method and Gibbs method estimated $\sigma^2(h)$, relatively large bias appeared and the RMSE was also relatively large, whereas the bias of $\sigma^2(h)$ under the HMC method remained relatively small.

The HMC algorithm is a sampling method that combines the physical properties of Hamiltonian systems with the Leapfrog technique to perform sampling. It is a global-update algorithm; it has relatively high computational efficiency and can handle solution problems for complex models (Lopilato et al., 2015; Nugroho & Morimoto, 2015; Pakman & Paninski, 2014). The HMC algorithm demonstrates its advantages: in the two-facet crossed design, under the noninformative prior condition, the HMC algorithm can accurately estimate $\sigma^2(h)$, whereas the estimation biases of both the M-H algorithm and Gibbs sampling are relatively large. This indicates that the HMC algorithm has strong robustness.

5.2 The turning point in the number of levels of facet h in the two-facet crossed design

In the two-facet crossed design, the estimation results under the noninformative prior condition may be related to the fact that the number of levels of facet h is 2. In subsequent research, as the number of levels of facet h increased, the estimation bias of the M-H algorithm and Gibbs sampling gradually decreased. In particular, after it increased to 4, their estimation results were closer to the estimation performance under the informative prior and empirical prior conditions. This indicates that, under the noninformative prior condition, the M-H algorithm and Gibbs sampling perform relatively poorly in estimating $\sigma^2(h)$, and this is related to the relatively small number of levels of facet h.

Moore (2010) examined the numbers of measurement-facet levels used in previous studies, and used the ratio of examinees to items (Ratio of persons to items) to represent the ratio of facet levels, obtaining a range from 0.13 to 120. Based on the ratios of facet levels in previous studies, Moore (2010) determined the number of measurement facets in the corresponding studies and found that the number of measurement-facet levels would have a certain influence on the estimation results for variance components. When the number of levels of facet h was 2, the estimates from the M-H algorithm and Gibbs sampling were unsatisfactory, whereas those from the HMC algorithm were satisfactory. However, when the number of levels of facet h was 4 or above, the estimates from the M-H algorithm and Gibbs sampling were comparable to those from the HMC algorithm. This further indicates that the HMC algorithm has strong robustness and can overcome the influence brought about by the number of levels of facet h. If the M-H algorithm and Gibbs sampling are to achieve estimation performance comparable to that of the HMC algorithm, then the number of levels of facet h in the two-facet crossed design must be at least 4. This suggests that, in the two-facet crossed design, a turning point appears when the number of levels of facet h is 4.

In most cases, the HMC algorithm shows estimation performance consistent with that of the M-H algorithm and Gibbs sampling, and all can estimate the variance components of the single-facet crossed design and the two-facet crossed design relatively accurately. However, the M-H algorithm and Gibbs sampling are easily affected by the number of levels of facet h, showing insufficient robustness. In summary, compared with Gibbs sampling and the M-H algorithm, the HMC algorithm performs better overall and is worthy of promotion.

5.3 Effects of prior information and missing-data proportion on variance-component estimation

Prior information affects the estimation of variance components, whereas the missing-data proportion has little influence on the estimation of variance components. In the single-facet crossed design, under the three prior-distribution conditions, the differences in variance-component estimation performance are relatively small, and the accuracy of variance-component estimation is relatively high under both the informative prior and empirical prior conditions. In the two-facet crossed design, the influence of the three types of prior information on the estimation results for variance components is relatively evident; here as well, the accuracy of variance-component estimation is better under the informative prior and empirical prior conditions, which is consistent with previous research findings (Lopilato et al., 2015).

In the single-facet crossed design and the two-facet crossed design, under the three missing-data proportion conditions, the differences in variance-component estimation performance are all very small. This further verifies that MCMC methods can handle missing data in GT relatively well, and that MCMC methods handle missing data in GT

showed strong robustness. The MCMC methods can handle missing data in GT analysis very effectively and display clear advantages; this is unmatched by other variance-component estimation methods.

From the perspective of the simulation-study runtime, compared with the one-facet crossed design, the two-facet crossed design required more time, indicating that the more complex the research design, the longer the computation time of the MCMC methods. Compared with the M-H algorithm and Gibbs sampling, the HMC algorithm required slightly longer runtime. This may be related to the complexity of the algorithm, but given that current computers operate relatively quickly, a small increase in runtime will not seriously affect the estimation results.

5.4 Limitations and Prospects

First, this study involved only two common measurement designs in GT research —the one-facet crossed design and the two-facet crossed design. Future research may continue to examine the performance of different sampling methods in estimating variance components under other measurement designs, such as nested designs and mixed designs (Vispoel et al., 2023). Second, missing data are relatively common in actual measurement processes. This study examined only the effect of completely random missingness on the accuracy of variance-component estimation. Future research may incorporate other missing-data mechanisms, such as random missingness and nonrandom missingness (Van Hooijdonk et al., 2022). Third, the simulation study fixed the number of students and items, and $h \geq 4$ was obtained under the two designs $p \times i$ and $p \times i \times h$. Future research may examine the effects of changing the number of students and items on variance-component estimation, and may also investigate the influence of the sample size of the h facet on variance-component estimation under more designs.

6 Conclusion

(1) A small amount of missing data has almost no effect on the three MCMC sampling methods, indicating that MCMC has relatively strong robustness for estimation with missing data.

(2) Overall, for variance-component estimation, estimations with informative priors and empirical priors are more accurate than estimations with non-informative priors.

(3) Only the HMC algorithm showed "cross-design" properties, with high estimation accuracy under both designs; however, the M-H algorithm and Gibbs sampling were constrained by $\sigma^2(h)$, and their estimation accuracy was not high.

(4) If the M-H algorithm and Gibbs sampling are to overcome the limitation of low estimation accuracy for $\sigma^2(h)$, it is necessary to set $h \geq 4$. For future researchers using MCMC methods, the HMC algorithm is more worthy of recommendation.

References

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L., Kim, S. Y., & Lee, W. C. (2022). Extended multivariate generalizability theory with complex design structures. *Educational and Psychological Measurement*, 82(4) 617-642.

Bourabee, N., & Sanzserna, J. M. (2017). Randomized Hamiltonian Monte Carlo. *Annals of Applied Probability An Official Journal of the Institute of Mathematical Statistics*, 27(4), 2159–2194.

Chaudhuri, S., Mondal, D., & Yin, T. (2017). Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 79*(1), 293–320.

Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal, 59*, 1122–1156.

Enders, C. K., Fairchild, A. J., & MacKinnon, D. P. (2013). A Bayesian approach for estimating mediation effects with missing data. *Multivariate Behavioral Research, 48*(3), 340–369.

Huang, C. (2017). *Bayesian statistics and its implementation in R.* Beijing: Tsinghua University Press.
[Huang Changquan. (2017). *Bayesian statistics and its implementation in R.* Beijing: Tsinghua University Press.]

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

Jackson, D. J. R., Michaelides, G., Dewberry, C., & Kim, Y. J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology, 101*(7), 976–994.

Jiang, Z., & Skorupski, W. (2018). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior Research Methods, 50*, 2193–2214.

Li, G. (2016). Influence factors of estimating of variance components and their variability for generalizability study in generalizability Theory. *Psychological Exploration, 36*(5), 458–463.
[Li Guangming. (2016). Factors influencing the estimation of variance components and their variability in G studies in generalizability theory. *Psychological Exploration, 36*(5), 458–463.]

Li, G. (2023). How many students and items are optimal for teaching level evaluation of college teachers? Evidence from generalizability theory and Lagrange multiplier. *Sustainability, 15*, 2.

Li, G., & Qin, Y. (2022). A new method for estimating the optimal sample size in generalizability theory based on evolutionary algorithm: Comparisons with three traditional methods. *Acta Psychologica Sinica, 54*(10), 1262–1276.
[Li Guangming, Qin Yue. (2022). A new method for estimating the optimal sample size in generalizability theory based on an evolutionary algorithm: Comparison with three traditional methods. *Acta Psychologica Sinica, 54*(10), 1262–1276.]

Li, G., Wang, X., & Pan, Y. (2022). Comparing of different methods for estimating variance components and their variabilities in generalizability theory based on two-facet designs. *Statistics and Decision, 38*(3), 50–54.
[Li Guangming, Wang Xingjun, Pan Yuxi. (2022). Comparison of methods for estimating variance components and their variability in two-facet designs in generalizability theory. *Statistics and Decision, 38*(3), 50–54.]

Li, G., & Wu, Y. (2023). Application of the Bookmark method: Setting standard for the ninth-grade mathematics achievement test in China. *Current Psychology, 42*, 28941–28952.

Liao, R. J. T. (2023). The use of generalizability theory in investigating the score dependability of classroom-based L2 reading assessment. *Language Testing, 40*(1), 86–106.

Liu, J., & Xia, Q. (2016). *Bayesian statistical method based on MCMC algorithm*. Beijing: Science Press.
[Liu Jinshan, Xia Qiang. (2016). *Bayesian statistical methods based on the MCMC algorithm*. Beijing: Science Press.]

Lopilato, A. C., Carter, N., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management: Official Journal of the Southern Management Association, 41*(2), 692–717.

Mao, X., Shin, D., & Brennan, R. (2005). *Estimating the variability of the estimated variance components and related statistics using the MCMC procedure: An exploratory study*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Medvedev, O. N., & Siegert, R. J. (2022). *Generalizability theory*. In Handbook of Assessment in Mindfulness Research (pp. 1–12). Cham: Springer International Publishing.

Medvedeva, A., Galli, G., Petrovskaya, A., & Medvedev O. N. (2022). A novel application of generalizability theory to evaluate the reliability of the recognition memory test. *Current Psychology, 42*(20), 17020–17032.

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution, 8*(3), 339–348.

Moore, J. L. (2010). *Estimating standard errors of estimated variance components in generalizability theory using bootstrap procedures*. (Doctor of Philosophy Doctor), University of Iowa.

Neal, R. M. (2010). *MCMC using Hamiltonian dynamics*. London: Chapman and Hall.

Nugroho, D. B., & Morimoto, T. (2015). Estimation of realized stochastic volatility models using Hamiltonian Monte Carlo-Based methods. *Computational Statistics, 30*(2), 491–516.

Pakman, A., & Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational & Graphical Statistics, 23*(2), 518–542.

Tong, Y., & Brennan, R. L. (2007). Bootstrap estimates of standard errors in generalizability theory. *Educational and Psychological Measurement, 67*(5), 804–817.

Vispoel, W. P., Hong, H., & Lee, H. (2023). Analyzing complete generalizability theory designs using structural equation models. *Applied Measurement in Education, 36*(4), 372–393.

Van Hooijdonk, M., Mainhard, T., Kroesbergen, E. H., & van Tartwijk, J. (2022). Examining the assessment of creativity with generalizability theory: An analysis of creative problem solving assessment tasks. *Thinking Skills and Creativity, 43*, 100994.

Wang, X., Li, G., Zhang, M., Jiang, X., Liang, Z., & Chu, X. (2016). A generalizability analysis of teaching level evaluation for college teachers. *Journal of Psychological Science, 39*(1), 90–96.

[Wang Xingjun, Li Guangming, Zhang Minqiang, Jiang Xin, Liang Zhengyan, Chu Xiaoyi. (2016). A generalizability-theory analysis of teaching level evaluation for college teachers. *Psychological Science, 39*(1), 90–96.]

Yang, Z., & Chang, L. (2003). *Generalizability theory and its applications.* Beijing, China: Educational Science Publishing House.

[Yang Zhiming, Zhang Lei. (2003). *Generalizability theory and its applications.* Beijing: Educational Science Publishing House.]

Zhang, C., Shahbaba, B., & Zhao, H. (2017). Hamiltonian Monte Carlo acceleration using surrogate functions with random bases. *Statistics and Computing, 27*(6), 1473–1490.

Zhao, X., Ren, Z., & Chen, A. (2018). Analysis and comparative study on the quality of mathematics, arts, and science test papers in the college entrance examination based on multivariate generalizability theory. *Mathematics Bulletin, 57*(1), 25–30

[Zhao Xuan, Ren Zichao, Chen Ang. (2018). Analysis and comparative study of the quality of mathematics, liberal-arts, and science test papers in the college entrance examination based on multivariate generalizability theory. *Mathematics Bulletin, 57*(1), 25–30.]

Zheng, X., Liu, Y., & Li, G. (2025). Using multivariate generalizability theory to test construct validity of the in-basket test and optimize evaluation plan. *Journal of Jiangxi Normal University (Natural Science Edition), 49*(6), 623–634.

[Zheng Xueqin, Liu Yaozhong, Li Guangming. (2025). Testing the construct validity of the in-basket test and optimizing the evaluation plan based on multivariate generalizability theory. *Journal of Jiangxi Normal University (Natural Science Edition), 49*(6), 623–634.]

Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12*(1), 63–70.

ChinaRxiv 中X

Comparing the accuracy of three sampling algorithms of MCMC method for estimating variance components in generalizability theory

LI Guangming

(School of Psychology, Center for Studies of Psychological Application, South China Normal University, Guangzhou 510631, China)

Abstract

Generalizability theory (GT), a modern psychometric framework, has been widely applied in practice. Its efficacy relies on the accurate estimation of variance components. While variance components are traditionally estimated using methods such as analysis of variance (ANOVA), Bootstrap, Jackknife, MINQUE, and restricted maximum likelihood (REML method), Bayesian Markov Chain Monte Carlo (MCMC) estimation has demonstrated advantages. However, the performance of MCMC depends on the chosen sampling algorithms. Local update algorithms, such as Metropolis-Hastings (M-H) algorithm and Gibbs sampling, are prone to random walk behaviors and slow convergence, particularly when conditional distributions are complex or parameters are highly correlated. Although the Hamiltonian Monte Carlo (HMC) algorithm effectively suppresses random walk behavior and handles high-dimensional, complex models efficiently, its application to variance component estimation within GT remains unexamined. To date, no published studies have compared M-H, Gibbs, and HMC algorithms in this context.

This study evaluated and compared the estimation accuracy of three MCMC sampling algorithms across two common GT research designs: the two-facet crossed ($p \times i$) design and the three-facet crossed ($p \times i \times h$) design. Using simulation methodologies, we manipulated three independent variables: (1) MCMC sampling algorithm (M-H, Gibbs, and HMC), (2) prior distributions (informative, non-informative, and empirical), and (3) missing data ratios (0%, 5%, and 10%). The dependent variables were the estimated variance components, Bias, and Root Mean Square Error (RMSE).

In simulation study, we generated data using R software, with 1000 batches of data generated for $p \times i$ and $p \times i \times h$. For the generated data, R software implemented variance component estimation for M-H, Gibbs, and HMC algorithms by calling the corresponding software package. In empirical research, a batch of data containing 1429 participants was used to validate the conclusions drawn from the simulation study, and it was found that the empirical research results were almost consistent with the simulation research results.

The research results indicated that: (1) low ratios of missing data (up to 10%) yielded negligible effects across all three algorithms, demonstrating the robust handling of missingness inherent to MCMC; (2) across all conditions,

ChinaRxiv 中X

informative and empirical priors yielded higher estimation accuracy than non-informative priors; (3) HMC uniquely exhibited "cross-design" stability, maintaining high accuracy across both the $(p \times i)$ and $(p \times i \times h)$ designs, whereas M-H and Gibbs sampling demonstrated structural limitations and lower accuracy; and (4) for M-H and Gibbs sampling to achieve acceptable accuracy in higher-dimensional spaces, a minimum of four levels for the third facet $(h \geq 4)$ was required. Consequently, we highly recommend that future GT researchers utilizing Bayesian MCMC frameworks adopt the HMC sampling algorithm.

Key words generalizability theory, estimating variance components, MCMC method, M-H algorithm, Gibbs sampling, HMC algorithm

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv. Machine translation. Verify with the original.