《心理学报》审稿意见与作者回应

题目：基于 MCMC 三种抽样方法的概化理论方差分量估计准确性比较

作者：黎光明

第一轮

审稿人 1 意见：

本研究围绕 MCMC 三种抽样方法（M-H、Gibbs、HMC）在概化理论（GT）方差分量估计中的准确性进行了系统比较，研究选题具有理论价值和实践意义。研究兼具模拟研究与实证分析，整体结构清晰，内容充实。但是本文在模拟研究的设计和分析，以及语言表述等方面仍有一些待改进之处。以下是具体的修改意见。

本研究围绕 MCMC 三种抽样方法（M-H、Gibbs、HMC）在概化理论（GT）方差分量估计中的准确性进行了系统比较，研究选题具有理论价值和实践意义。研究兼具模拟研究与实证分析，整体结构清晰，内容充实。但是本文在模拟研究的设计和分析，以及语言表述等方面仍有一些待改进之处。以下是具体的修改意见。

1. 模拟研究

围绕本文的模拟研究设计和分析，有两个关键问题需要指出。

首先，“M-H 算法通过在 R 软件中调用 BRugs 包与 OpenBUGS 软件链接来实现，Gibbs 抽样是在 R 软件中调用 R2jags 包与 JAGS 软件链接实现”这样的表述不太合适，因为这两种软件在共轭先验下都默认采用 Gibbs 抽样。BUGS 和 JAGS 的核心共性之一在于智能化的算法选择，意味着并不会固定使用单一的 MCMC 算法。如果某个参数的条件后验分布有已知的标准解析形式，软件会为其分配 Gibbs 采样器；对于非共轭的、条件后验分布形式未知或复杂的参数，软件会为其分配基于 M-H 算法的采样器。具体到本研究，由于方差分量都设置了共轭先验逆伽马分布，那么合理预期 BUGS 和 JAGS 都会选择 Gibbs 采样（实际上，即使单个参数是共轭先验，软件也可能引入多变量的 M-H 采样器来对一组高相关的参数进行联合更新从而提升采样效率）。一般而言，研究者并不好手动控制 BUGS 和 JAGS 采用何种采样器，只能事后检查软件在每个节点所使用的采样器类型。因此，如果一定要比较 M-H 和 Gibbs 采样，建议作者采用能够手动设置抽样方法的软件（例如 Karunarasan 等(2021)采用的 MLwiN），或者考虑支持贝叶斯估计的更加灵活和综合的 R 包，例如 NIMBLE

(de Valpine et al., 2024)。当然，研究者也可以将模拟研究的目的由比较三种抽样方法改为比较三种估计软件（BUGS, JAGS, Stan）的表现，但如此一来，对于本研究的理论价值以及文中多处相关表述就需要作重新考量。

de Valpine P, Paciorek C, Turek D, Michaud N, Anderson-Bergman C, Obermeyer F, Wehrhahn Cortes C, Rodrìguez A, Temple Lang D, Paganin S. (2024). NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling. doi:10.5281/zenodo.1211190, R package version 1.3.0, https://cran.r-project.org/package=nimble.

Karunarasan, D., Sooriyarachchi, R., & Pinto, V. (2021). A comparison of Bayesian Markov chain Monte Carlo methods in a multilevel scenario. Communications in Statistics - Simulation and Computation, 52(10), 4756 – 4772. https://doi.org/10.1080/03610918.2021.1967985

其次，模拟研究在两种设计下都仅考虑了一种 p×i 的组合，未见讨论学生人数和题量对方差分量估计准确性的影响，这会限制模拟研究所得结论的可推广性。尤其作者在摘要和正文中都给出了若采用 M-H 算法和 Gibbs 抽样估计双侧面交叉设计的方差分量须规定 h ≥ 4 的指导建议，而并未说明这一建议是基于 100 人×20 题这一组合的条件下，有失严谨。建议增加条件探索不同 p×i 的组合下如何选择 h 侧面水平数，或者至少在总结模拟研究时点出 h ≥ 4 建议的适用条件，并在讨论模拟研究的局限性时增加这一点。

关于模拟研究部分，还有一些相对小的问题需要作者注意修改，包括但不限于：

第 2 和 3 部分一级标题中的“p×i 设计”和“p×i×h 设计”最好分别改为“单侧面交叉设计（p×i）”和“双侧面交叉设计（p×i×h）”，另外建议将这两种设计的简介由 2.1 和 3.1 移到模拟研究前的第 1 部分;

虽然 3.2 中介绍了经验性先验，但 2.2 中没有介绍经验性先验具体如何设定；

介绍比较标准时公式逻辑不清：RMSE 中的 θ 应当是 1000 次估计结果中的单次估计值，但上一行的 Bias 公式中的 θ 却是对应 2.2（3）的文字部分中的“1000 次估计结果的平均值”；

2.4 中“需要模拟正态分布的 z_p、z_i 和 z_pi”，需说明是标准正态分布；

2.5 和 3.5 的估计结果都没有报告每种 MCMC 算法的收敛诊断情况；

表 2 和表 4 中的“VC”直到后文介绍表 7 和表 8 才给了中文解释，建议涉及了 VC 的每张表都加上注释，给出 VC 的中英文全称。

2. 其他表述问题

实证数据的来源没有说明：是作者团队收集的数据还是已有研究的数据？

实证分析部分“对 1429 名被试的数据先重复 5000 次估计方差分量的平均数，将其视为‘参数’”，请援引采用此做法的文献，在文中简要说明此做法的合理性；另外既然得到的是参数“真值”而不是真值，请作者斟酌此时算出的比较标准是否直接称作 Bias 和 RMSE，又该如何解释它们。

文中多处术语的全称或简称表述失当。例如：第 1 部分第 4 自然段先给出了“M-H 算法”后才给出“Metropolis-Hasting(M-H)算法”，“HMC 算法(Hamiltonian Monte Carlo)”应当是“HMC (Hamiltonian Monte Carlo) 算法”；2.2 中“有信息先验(with informative priors, MCMC inf)”的括号里全称漏了“MCMC”；3.5.1 出现了“均方根误差”和“方根误差”，应统一用 RMSE 代指。建议仔细检查并修改。

文中还有几处语句表述不通顺，例如：第 1 部分第 4 自然段“二是 MCMC 三种抽样方法应用于 GT 中的性能如何？尚需比较才能获知”应当删去“？”；第 2 部分第 1 自然段“探讨三种抽样方法的估计表现和方差分量的影响因素”建议改成“探讨三种抽样方法估计方差分量的表现及其影响因素”。建议通读全文，仔细检查并修改。

意见 1：

1. 模拟研究

围绕本文的模拟研究设计和分析，有两个关键问题需要指出。

首先，“M-H 算法通过在 R 软件中调用 BRugs 包与 OpenBUGS 软件链接来实现，Gibbs 抽样是在 R 软件中调用 R2jags 包与 JAGS 软件链接实现”这样的表述不太合适，因为这两种软件在共轭先验下都默认采用 Gibbs 抽样。BUGS 和 JAGS 的核心共性之一在于智能化的算法选择，意味着并不会固定使用单一的 MCMC 算法。如果某个参数的条件后验分布有已知的标准解析形式，软件会为其分配 Gibbs 采样器；对于非共轭的、条件后验分布形式未知或复杂的参数，软件会为其分配基于 M-H 算法的采样器。具体到本研究，由于方差分量都设置了共轭先验逆伽马分布，那么合理预期 BUGS 和 JAGS 都会选择 Gibbs 采样（实际上，即使单个参数是共轭先验，软件也可能引入多变量的 M-H 采样器来对一组高相关的参数进行联合更新从而提升采样效率）。一般而言，研究者并不好手动控制 BUGS 和 JAGS 采用何种采样器，只能事后检查软件在每个节点所使用的采样器类型。因此，如果一定要比较 M-H 和 Gibbs 采样，建议作者采用能够手动设置抽样方法的软件（例如 Karunarasan 等

(2021)采用的 MLwiN），或者考虑支持贝叶斯估计的更加灵活和综合的 R 包，例如 NIMBLE (de Valpine et al., 2024)。当然，研究者也可以将模拟研究的目的由比较三种抽样方法改为比较三种估计软件（BUGS, JAGS, Stan）的表现，但如此一来，对于本研究的理论价值以及文中多处相关表述就需要作做新考量。

de Valpine P, Paciorek C, Turek D, Michaud N, Anderson-Bergman C, Obermeyer F, Wehrhahn Cortes C, Rodrìguez A, Temple Lang D, Paganin S. (2024). NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling. doi:10.5281/zenodo.1211190, R package version 1.3.0, https://cran.r-project.org/package=nimble.

Karunarasan, D., Sooriyarachchi, R., & Pinto, V. (2021). A comparison of Bayesian Markov chain Monte Carlo methods in a multilevel scenario. Communications in Statistics - Simulation and Computation, 52(10), 4756 – 4772. https://doi.org/10.1080/03610918.2021.1967985

回应：感谢专家的意见。本研究者是通过手动控制 BRugs、BUGS 和 JAGS 采用何种采样器，其中前者是为了实现 M-H 算法，后两者是为了实现 Gibbs 抽样，现附上附录代码（请参见本修改稿中的“附录程序代码”，英文摘要之后），说明如下：

第一，对于 p×i 设计。通过调用 BRugs 软件，实现 M-H 算法的代码如下：modelEnable(factory = 'adaptive metropolis NLLS')（请参见本修改稿中的附录程序代码：第 34 页，第 20 行；第 35 页，第 26 行，都已涂成黄色高亮，方便找到），其意指调用一个函数调配置命令，涉及模型选择或优化算法的启用，其中 “adaptive metropolis NLLS”是一种优化方法或采样策略的名称。Metropolis 算法的英文全称是 Metropolis Algorithm，它是 Markov Chain Monte Carlo (MCMC) 方法的基础，常用于从复杂概率分布中采样；其扩展版本 Metropolis–Hastings algorithm 通常用于贝叶斯优化等场景。非线性最小二乘（NLLS）的英文全称是 Non-linear Least Squares，它是一种参数估计技术，通过最小化观测值与模型预测值之间残差的平方和来拟合非线性模型，广泛应用于曲线拟合和系统识别等领域。因此，照此看来，本研究的 M-H 算法采用的是 adaptive metropolis NLLS，可认为 M-H 算法通过在 R 软件中调用 BRugs 包链接来实现。

然而，通过 R 软件调用 R2jags 包与 JAGS 软件链接实现 Gibbs 抽样的代码如下：mcmc.result<-jags.parallel(data=data,inits=NULL,parameters.to.save =parameters,model.file=mcmc.model,n.chains=3,n.iter=10000,n.thin=1,n.cluster=3)（请参见本修改稿中的附录程序代码：第 37 页，第 6 行；第 38 页，第 17 行，都已涂成黄色高亮，

方便找到），这是用于行执行 JAGS（Just Another Gibbs Sampler）模型的函数调用，常用于贝叶斯统计推断。因为使用了 JAGS，并指令了 Gibbs 抽样，所以这里就是 Gibbs 抽样。 因此，照此看来，本研究的 Gibbs 算法采用的是 Just Another Gibbs Sampler，可认为 Gibbs 抽样是在 R 软件中调用 R2jags 包与 JAGS 软件链接实现。

第二，对于 p×i×h 设计。实现 M-H 算法的代码与 p×i 设计基本一致，也是通过调用 BRugs 软件，实现 M-H 算法，其代码也是：modelEnable(factory = 'adaptive metropolis NLLS')（请参见本修改稿中的附录程序代码：第 42 页，第 28 行；第 44 页，第 15 行，都已涂成黄色高亮，方便找到）。

然而，Gibbs 抽样是通过 R 软件调用 R2OpenBUGS 与 BUGS 软件链接实现的，其代码如下：mcmc_inf=bugs(data,inits=inits,parameters,model.file= "pih_model.txt", n.chains=3,DIC=FALSE,n.iter=10000,n.burnin=5000,n.thin=1, working.directory=NULL,clearWD=FALSE,codaPkg=TRUE,debug=FALSE)（请参见本修改稿中的附录程序代码：第 45 页，第 24 行；第 47 页，第 14 行，都已涂成黄色高亮，方便找到）。这里的 BUGS 函数也是使用默认的 Gibbs 抽样。

赞成审稿人的观点，原文表述的确出现了错误，现修改为：

2.3 数据模拟与估计

对于 p×i 设计，M-H 算法通过 R 软件调用 BRugs 包来实现；Gibbs 抽样通过 R 软件调用 R2jags 包与 JAGS 软件链接实现；HMC 算法通过 R 软件调用 Rstan 包来实现。

3.3 数据模拟与估计

对于 p×i×h 设计，M-H 算法通过 R 软件调用 BRugs 包来实现；Gibbs 抽样通过 R 软件调用 R2OpenBUGS 与 BUGS 软件链接实现；HMC 算法通过 R 软件调用 Rstan 包来实现。

修改稿中所有修改之处均用红色进行了标识，凡是红色之处均有一定修改。

意见 2：

其次，模拟研究在两种设计下都仅考虑了一种 p×i 的组合，未见讨论学生人数和题量对方差分量估计准确性的影响，这会限制模拟研究所得结论的可推广性。尤其作者在摘要和正文中都给出了若采用 M-H 算法和 Gibbs 抽样估计双侧面交叉设计的方差分量须规定 h ≥ 4 的指导建议，而并未说明这一建议是基于 100 人×20 题这一组合的条件下，有失严谨。建议增加条件探索不同 p×i 的组合下如何选择 h 侧面水平数，或者至少在总结模拟研究时

点出 h ≥ 4 建议的适用条件，并在讨论模拟研究的局限性时增加这一点。

回应：感谢专家的意见。得出 h ≥ 4 是基于本研究获得的，其基本条件是样本量固定为：p 的水平数是 100，i 的水平数是 20，已在局限中增加了第三点说明，如下：

其三，模拟研究固定学生人数和题量，未探讨改变样本量对方差分量估计的影响，且 $h \geq 4$ 是基于 p×i 和 p×i×h 两种设计之下所获得的。

意见 3：

关于模拟研究部分，还有一些相对小的问题需要作者注意修改，包括但不限于：

第 2 和 3 部分一级标题中的“p×i 设计”和“p×i×h 设计”最好分别改为“单侧面交叉设计（p×i）”和“双侧面交叉设计（p×i×h）”，另外建议将这两种设计的简介由 2.1 和 3.1 移到模拟研究前的第 1 部分;

回应：感谢专家的意见。第 2 和 3 部分一级标题已更改为“单侧面交叉设计(p×i)模拟研究”和“3 双侧面交叉设计(p×i×h)模拟研究”。已将 2.1 和 3.1 移到模拟研究前的第 1 部分，请见修改后的文档。

意见 4：

虽然 3.2 中介绍了经验性先验，但 2.2 中没有介绍经验性先验具体如何设定；

回应：感谢专家的意见。在原稿 2.2 中（修改 2.1 中），其单侧面交叉设计 p×i 的经验性先验具体如下：

根据贝叶斯统计学中共轭先验分布的性质(黄长全, 2017)，将先验信息的目标分布设为逆伽马分布(IGamma)。参考前人研究对先验信息的设定(Lopilato et al., 2015; Mao et al., 2005)。**错误!未找到引用源。**呈现了三种先验信息的具体设定情况。

表 1 单侧面交叉设计 p×i 的三种先验信息

先验信息	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
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)

意见 5：

介绍比较标准时公式逻辑不清：RMSE 中的 θ 应当是 1000 次估计结果中的单次估计值，但上一行的 Bias 公式中的 θ 却是对应 2.2(3)的文字部分中的“1000 次估计结果的平均值”；

回应：感谢专家的意见。已修正为如下：

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

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

在公式（1）和公式（2）中，$\hat{\theta}$表示每次模拟统计量的估计值，即方差分量的估计值；θ表示统计量的真值，即方差分量的真值；n 表示每种条件下的模拟次数。

意见 6：

2.4 中“需要模拟正态分布的 z_p、z_i 和 z_pi”，需说明是标准正态分布；

回应：感谢专家的意见。已修改为标准正态分布，如下：

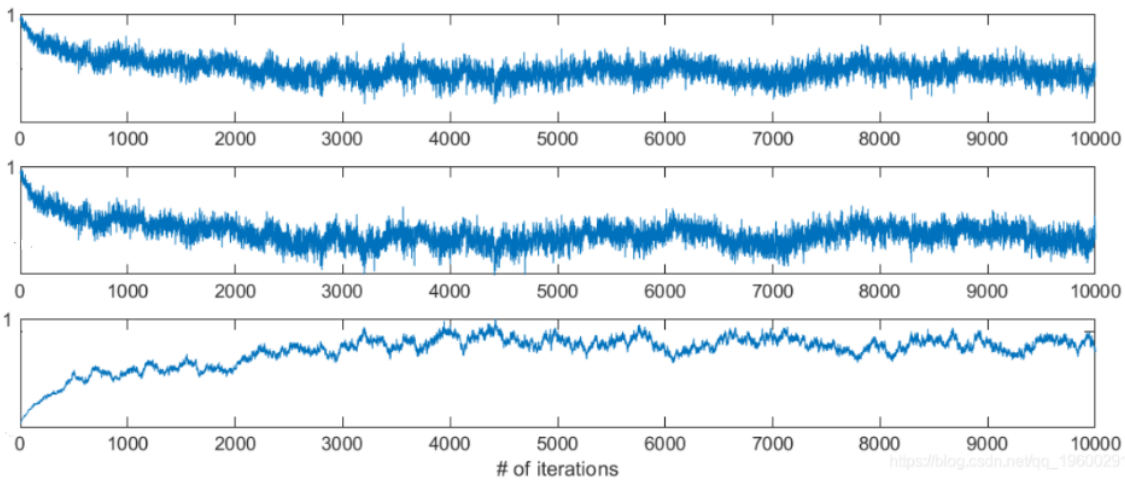
需要模拟标准正态分布的z_p、z_i和z_{pi}，分别与σ_p、σ_i和σ_{pi}相乘。

意见 7：

2.5 和 3.5 的估计结果都没有报告每种 MCMC 算法的收敛诊断情况；

回应：感谢专家的意见。对原稿的 2.5 和 3.5 的估计结果说明了 MCMC 算法的收敛诊断情况，如下：

根据问题和设置，MCMC 在序列接近目标分布之前可能需要进行多次迭代。我们以如 p×i 设计为例，$\sigma^2(p)$、$\sigma^2(i)$和$\sigma^2(pi)$的历史迭代图如下图所示。



从上图显示，三个方差分量估计值在 5000 次附近快速达到稳定状态。稳定性之前的这些迭代称为“老化”。一共运行 10000 次迭代，从第 5001 开始至 10000 次，每条 Markov 链的值都为估计值。

为了表述 MCMC 诊断情况，在论文中加了一些表述，如下：

2.4 估计结果

对于 p×i 设计，通过对历史迭代图发现， 模型通过 R 软件进行抽样，最终达到收敛。通过观察历史迭代图，设定预热次数为 5000 就能达到平稳分布，因此从第 5001 开始至 10000 次，每条 Markov 链的值都为估计值，对三条 Markov 链求平均，即为每个统计的估计值。

3.4 估计结果

3.4.1 不同缺失比例下的估计结果

对于 p×i×h 设计，通过对历史迭代图发现， 模型通过 R 软件进行抽样，最终达到收敛。通过观察历史迭代图，设定预热次数为 5000 就能达到平稳分布，因此从第 5001 开始至 10000 次，每条 Markov 链的值都为估计值，对三条 Markov 链求平均，即为每个统计的估计值。

意见 8：

表 2 和表 4 中的“VC”直到后文介绍表 7 和表 8 才给了中文解释，建议涉及了 VC 的每张表都加上注释，给出 VC 的中英文全称。

回应：感谢专家的意见。在每张表的说明中，都已给出了 VC 是方差分量的说明，请参见修改后的稿件。

意见 9：

2. 其他表述问题

实证数据的来源没有说明：是作者团队收集的数据还是已有研究的数据？

回应：感谢专家的意见。是作者团队收集的数据，已在正文中加入了对于数据的说明，如下：

采用 Zung(1965)编制的抑郁自评量表(Self-rating Depression Scale, SDS)对学生进行测量。SDS 包含 20 个反映抑郁主观感受的项目，其中，10 项为正向评分题，另 10 项为反向评分题。选项如“我觉得闷闷不乐，情绪低沉”“我觉得一天之中早晨最好”等，按症状出现的频度采用四级计分。1 表示“没有或很少时间”，2 表示“小部分时间”，3 表示“相当多时间”，4 表示“绝大部分或全部时间”。

数据来源于中国南方某城市随机取样的 1545 名学生，对他们进行了为期 6 个月的追踪测试，共进行了三次纸笔测试。三次测试均存在被试流失，其原因可能为：(1) 测试当天缺勤；(2)题目错填或漏填；(3)转学或休学。剔除未完成三次测试的被试，最后保留被试 1429 名，其中，男性青少年 636 名，女性青少年 793 名，平均年龄 14.88 岁(标准差 1.80)。该量表在本研究中的内部一致性表现良好(T1 Cronbach's α = 0.902; T2 Cronbach's α = 0.906, T3 Cronbach's α = 0.922)。

意见 10：

实证分析部分“对 1429 名被试的数据先重复 5000 次估计方差分量的平均数，将其视为‘参数’”，请援引采用此做法的文献，在文中简要说明此做法的合理性；另外既然得到的是参数“真值”而不是真值，请作者斟酌此时算出的比较标准是否直接称作 Bias 和 RMSE，又该如何解释它们。

回应：感谢专家的意见。我们依据下列文献的做法，使用“经验参数”(empirical parameters)代替真值作为参照标准。经验参数的产生是依据与正式模拟相同的数据生成方法模拟 5000 批数据矩阵，并使用传统 GT 方法计算这 5000 批数据的各方差分量等参数作为经验参数。具体来说，对于某一个方差分量估计值而言，5000 批数据下的 5000 个估计值的均值就作为该方差分量的经验参数。

Chien, Y. M. (2008). An investigation of testlet-based item response models with a random facets design in generalizability theory. Unpublished Doctoral dissertation，The University of Iowa．

Chien, Y. M. (2009). *An investigation of testlet-based item response models with a random facets design in Generalizability Theory*. Dissertation Abstracts International. Section A: Humanities and Social Sciences(No.2A), 543.

Tong, Y., & Brennan, R. L. (2006). *Bootstrap techniques for estimating in generalizability theory* (CASMA Research Report No 15). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. Available from http://www.education.uiowa.edu/casma.

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

胡小甜, 张敏强, 田文娜, 梁淑仪, 张楠楠, 黄牧蕙. (2013). 不同参数分布形态下 GIRM 方法和传统 GT 方法的对比研究. *心理学探新, 33*(3), 246–251.

胡小甜, 张敏强, 郭凯茵, 黎光明. (2014). GIRM 方法与传统 GT 方法的比较. *统计与决策, 399*(3), 89–92.

意见 11：

文中多处术语的全称或简称表述失当。例如：第 1 部分第 4 自然段先给出了“M-H 算法”后才给出“Metropolis-Hasting(M-H)算法”，“HMC 算法(Hamiltonian Monte Carlo)”应当是“HMC (Hamiltonian Monte Carlo) 算法”；

回应：感谢专家的意见。已更正。非常感谢！

意见 12：

2.2 中“有信息先验(with informative priors, MCMC inf)”的括号里全称漏了“MCMC”；

回应：感谢专家的意见。已更正。非常感谢！

意见 13：

3.5.1 出现了“均方根误差”和“方根误差”，应统一用 RMSE 代指。建议仔细检查并修改。

回应：感谢专家的意见。已统一用 RMSE 代指，已修正。非常感谢！

意见 14：

文中还有几处语句表述不通顺，例如：第 1 部分第 4 自然段“二是 MCMC 三种抽样方法应用于 GT 中的性能如何？尚需比较才能获知”应当删去“？”；第 2 部分第 1 自然段“探讨三种抽样方法的估计表现和方差分量的影响因素”建议改成“探讨三种抽样方法估计方差分量的表现及其影响因素”。建议通读全文，仔细检查并修改。

回应：感谢专家的意见。已修改为：二是 MCMC 三种抽样方法应用于 GT 中的性能如何？迄今为止仍没有学者进行过探究。已修改为：探讨三种抽样方法估计方差分量的表现及其影响因素。虚心接受专家的观点，已通读全文，仔细检查并修改，所有修改之处都已用红色标识出来了。非常感谢。

审稿人 2 意见：

主要问题：

单面交叉设计和双面交叉设计的数据往往比较好，方差分量估计时不太容易出现问题，若使用嵌套设计作为研究对象，其结果不知是否类似？根据以往研究，非交叉设计的方差分量估计有可能出现负数这种违背常理的结果，作者可以在这个主题上展开讨论吗？

细节问题：

1）句式冗长：

原文 (P5, 2.2 研究设计): “参考以往研究对单侧面交叉设计 p×ip×i 的研究设定(Mao et al., 2005)，将方差分量...”

问题与分析：“对...的研究设定”中“研究”一词重复出现，读起来不够流畅。

修改建议：可精简为“参考以往关于单侧面交叉设计 p×ip×i 的设定(Mao et al., 2005)”，或“参考单侧面交叉设计 p×ip×i 的经典研究设定(Mao et al., 2005)”。

2）语言表达：

问题：英文摘要中 Generalization theory(GT) is amodern psychological and educational measurement theory, which has been widelyapplied in practice. 这个句子 which 的用法不符合英语学术写作中“非限制性定语从句”的规范，显得非常中式英语。建议重写，例如：Generalizability Theory (GT), a modern psychometric theory, has beenwidely applied in practice.

意见 1：

主要问题：

单面交叉设计和双面交叉设计的数据往往比较好，方差分量估计时不太容易出现问题，若使用嵌套设计作为研究对象，其结果不知是否类似？根据以往研究，非交叉设计的方差分量估计有可能出现负数这种违背常理的结果，作者可以在这个主题上展开讨论吗？

回应：感谢专家的意见。对于嵌套设计或非交叉设计，本研究在局限中已作出如下说明：其一，仅涉及到了 GT 研究中的两种常见测验设计——单侧面交叉设计和双侧面交叉设计，未来研究可以继续探讨不同抽样方法在其他的测验设计中对方差分量的估计表现，如嵌套设计和混合设计等(Vispoel et al., 2023)。

意见 2：

细节问题：

1）句式冗长：

原文 (P5, 2.2 研究设计):“参考以往研究对单侧面交叉设计 p×i 的研究设定(Mao et al., 2005)，将方差分量...”

问题与分析：“对...的研究设定”中“研究”一词重复出现，读起来不够流畅。

修改建议：可精简为“参考以往关于单侧面交叉设计 p×i 的设定(Mao et al., 2005)”，或“参考单侧面交叉设计 p×i 的经典研究设定(Mao et al., 2005)”。

回应：感谢专家的意见。已修改，如下：

2.1 研究设计

参考以往关于单侧面交叉设计 p×i 的设定(Mao et al., 2005)，

3.1 研究设计

参考以往关于双侧面交叉设计 p×i×h 的设定(Tong & Brennan, 2007; 王幸君等, 2016)

非常感谢！

意见 3：

2）语言表达：

问题：英文摘要中 Generalization theory(GT) is a modern psychological and educational measurement theory, which has been widely applied in practice. 这个句子 which 的用法不符合英语学术写作中“非限制性定语从句”的规范，显得非常中式英语。建议重写，例如：Generalizability Theory (GT), a modern psychometric theory, has been widely applied in practice.

回应：感谢专家的意见。已对英文摘要进行认真且仔细地核对，把有问题的地方都进行了红色标识，敬请查看修改后的论文。非常感谢！

第二轮

审稿人 1 意见：

作者较好回应了上一轮审稿提出的问题，文章质量得到较大提升。建议在补充线上公开材料中提供实证数据分析的代码，并提供分析所用数据。另外，请仔细检查文章的语言表述，以及参考文献的格式，经过上一轮修改完善后，语言表述上有些地方存在重复，请修改完善。

意见 1：

作者较好回应了上一轮审稿提出的问题，文章质量得到较大提升。建议在补充线上公开材料中提供实证数据分析的代码，并提供分析所用数据。另外，请仔细检查文章的语言表述，以及参考文献的格式，经过上一轮修改完善后，语言表述上有些地方存在重复，请修改完善。

回应：感谢专家的意见。针对上述修改意见，我们做了如下处理：

第一，在线上公开材料中，我们补充了实证数据分析的程序代码，并提供了实证分析所使用的 1429 个数据，详见线上公开材料（附件）。

第二，遵照 APA 第七版格式，对本研究的参考文献的格式进行了严格地核对，并加以修订，所有修订之处皆以“蓝色”进行了标注。

第三，仔细检查了文章的语言表述，特别是对语言表述上存在重复的地方进行了修改和完善，在正文中所有修订之处皆以“蓝色进行了标注，请详见本文的这份修改稿。

编委意见：

这篇论文围绕 MCMC 三种抽样方法（M-H、Gibbs、HMC）在概化理论（GT）方差分量估计中的准确性进行了系统比较，研究选题具有理论价值和实践意义。经过作者和审稿人互动式审改，论文达到发表要求，建议发表。

主编意见：

同意发表。