·统计、测量与方法·

处理缺失作答和随机猜测的认知诊断模型开发及其应用研究*

李潇沛[1] 彭思韦**[2] 王 琴**[1] 蔡 艳[1]
（[1]江西师范大学心理学院，南昌，330022）（[2]浙江师范大学心理学院，金华，321004）

摘 要 在实际测验中，被试的异常作答，尤其是缺失作答和随机猜测，往往会导致参数估计的偏差并且损害测验结果的准确性和公平性。然而，目前在认知诊断领域中，针对异常作答的建模研究仍然十分有限。针对这一现状，本研究首次尝试将项目反应树模型与认知诊断模型联合建模，开发出一种新型的认知诊断模型——IRTree–LCDM，该模型能够同时考虑缺失作答和随机猜测的影响。为评估新模型的性能及其在实证数据中的效果，研究采用 Monte Carlo 模拟实验与真实数据分析相结合的研究方法。模拟研究结果表明，新开发的 IRTree–LCDM 在各种实验条件下的参数估计精度表现良好。同时，与传统认知诊断模型相比，IRTree–LCDM 的判准率更为精准，对被试单个属性的判准率均值超过 .946，模式判准率均值达到 .783。此外，IRTree–LCDM 在实证数据中能够更好地拟合真实数据，且对被试的属性掌握模式的估计也更加合理。这些结果表明，IRTree–LCDM 在处理异常作答方面具有显著的价值和意义。

关键词 认知诊断 项目反应树模型 项目反应理论 缺失作答 随机猜测

1 引言

在心理与教育测量领域，测验施测者通过分析被试的作答反应来评估其心理特质或能力。这一过程传统上依赖于宏观测量技术，如项目反应理论（item response theory, IRT），该方法通过将测量特质和项目特性参数化，利用作答数据进行估计分析，从而评估被试的总体特质水平。然而，新一代心理与教育测量技术中的认知诊断评估（cognitive diagnosis assessment, CDA）方法则不局限于对被试宏观特质的评估。CDA 专注于分析个体在微观层面的认知过程、加工技能或知识结构（涂冬波等，2019）。与传统方法不同，CDA 通过对测验结果的深入分析，能够提供被试在具体知识点、技能掌握和认知能力上的详细诊断。这种方法不仅揭示了被试的知识和技能掌握情况，还为个性化教育和精准干预提供了宝贵的信息。作为新一代测量技术的核心，CDA 在准确测量和评估被试的微观认知过程、加工技能和知识结构方面发挥了至关重要的作用，它通过提供更细致的反馈和诊断，帮助教育者更好地了解学生的学习需求，从而进行个性化教学和有效的教育干预。

在实施心理与教育测验时，研究者们期望通过一系列精心设计的问题来揭示受测者的心理状态、能力特质以及行为倾向。然而，实际操作过程中，测验施测者们常常面临一个复杂的挑战，即受测者的作答行为有时并不能准确地反映其真实的心理状态或特质（刘玥，刘红云，2021）。这种现象背后的原因复杂多样，涉及多个层面，包括测试环境的具体情况、个体的主观能动性、答题时的心理活动过程、作答的动机水平等因素。越来越多的研究发现，被试在测验过程中会出现一些除了正常作答之外的异常反应行为，例如缺失作答（miss response; Holman & Glas, 2005）、猜测作答（guess response; Wise & DeMars, 2006）、项目预知（item preknowledge; McLeod et al., 2003）及机械作答（mechanical response; Kim & Moses, 2018） 等。

* 本研究得到国家自然科学基金项目（62467002, 32160203, 62167004, 32300942）的资助。
** 通讯作者：彭思韦，E–mail: siweipeng@zjnu.edu.cn; 王琴，E–mail: 005490@jxnu.edu.cn
DOI:10.16719/j.cnki.1671–6981.20260119

研究表明，异常作答的出现不仅会导致数据分析过程中产生误差,还会引发模型参数估计结果的偏差。这种偏差可能会削弱测验的信度和效度，从而影响整体的测量精度和测验结果的解释。最终，这些偏差可能使得研究者无法准确评估和诊断被试的真实心理特质或能力，进而影响后续的教育干预和决策（Glas & Pimentel, 2008; Hsu et al., 2020; Köhler et al., 2015; Kuha et al., 2018）。

在众多异常作答行为中，研究者们特别关注两种最为常见且影响重大的行为类型：缺失作答和随机猜测。在心理与教育测验中，缺失作答和随机猜测的存在非常常见（Cao & Stokes, 2008; Liu, 2021）。缺失作答通常发生在被试无法或选择不作答某些试题的情况下，在长时间测试、复杂问题情境下或被试对测试内容不熟悉时尤为常见。可能的原因包括时间限制、答题策略以及对题目理解的不足等（Huang, 2020; Liu, 2021）。随机猜测一般出现在选择题中，被试在低作答动机、作答时间不足或认知疲劳的情况下，通过随意选择答案以尽快完成测试（Jin et al., 2022; Lu et al., 2020; Rios, 2022; Ulitzsch et al., 2020b）。这往往与所测量的能力或特质无关，不能显现出被试的能力或特质的真实水平（Cao & Stokes, 2008; Goegebeur et al., 2008; Jin et al., 2022）。缺失作答和随机猜测这两种异常行为不仅在测试中频繁出现，而且对测验结果产生了深远的负面影响。缺失作答的直接后果是导致数据的不完整性，削弱了测验的信度和效度，使测验无法全面反映被试的真实能力和特质。随机猜测则引入了显著的噪声和误差，可能导致结果偏离真实值，使模型参数的估计偏差增大，从而降低测验的整体准确性和区分度。值得一提的是，在实际测验过程中，随机猜测和缺失作答这两种异常行为有时会同时出现，并共同影响测验结果。这种双重异常行为不仅增加了数据分析和模型拟合的复杂性，还使得对被试能力和心理特质的评估更加不确定和不稳定。具体来说，随机猜测可能导致被试得分虚高，而缺失数据可能导致得分偏低或数据不完整。这两种行为的相互作用使得测验结果难以准确反映被试的真实能力水平。

目前，在认知诊断领域中，针对缺失作答和随机猜测的研究相对匮乏。已有的研究大多集中于分析单一的异常行为，而较少探讨多个异常反应并存的情形。此外，现有的研究中，缺乏对被试多种异常作答倾向进行参数化估计的认知诊断模型。基于以上背景，本文旨在提出一个能够同时考虑缺失作答和随机猜测两种异常反应的认知诊断模型（即 IRTree-LCDM）。值得特别说明的是，项目未作答是心理与教育测量中缺失数据形成的主要原因。被试对某项目未作答的原因有很多，可能是疏忽忘记作答该题，可能是不会作答该题，也可能是在规定时间内没有解出该题答案，而本研究聚焦的缺失作答是指受到潜在缺失作答倾向的影响而产生的缺失作答，是一种不可忽视的缺失作答。模型通过估计被试的缺失反应倾向、随机猜测倾向及相关项目参数，以减少异常作答所导致的诊断误差。此模型的应用有望降低因异常反应引起的估计偏差，提高对被试认知状态的诊断准确性，并进一步丰富认知诊断领域中关于异常反应的研究。

在下文中，本文将首先介绍 IRTree-LCDM 建模的基础知识，随后详细阐述新模型的建模思路和具体构建过程。接着，通过模拟研究考察新模型的参数估计精度，并将其与传统的认知诊断模型进行对比。最后，本文将新模型应用于真实数据，探讨其在实际数据分析中的表现和应用效果。

2 建模基础

2.1 对数线性认知诊断模型（log-linear CDM, LCDM)

对数线性认知诊断模型（log-linear CDM, LCDM）（Henson et al., 2009）包含了截距效应、属性主效应和属性间的交互效应，其函数表达式如公式（1）和公式（2）：

$$P(Y_{ni}=1|\alpha_n)=\frac{exp\left(\lambda_{i,0}+\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_n,\boldsymbol{q}_i)\right)}{1+exp\left(\lambda_{i,0}+\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_n,\boldsymbol{q}_i)\right)}, \quad (1)$$

$$\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_n,\boldsymbol{q}_i)=\sum_{k=1}^{K}\lambda_{i,k}\alpha_{nk}q_{ik}+\sum_{k=1}^{K-1}\sum_{k=k+1}^{K}\lambda_{i,kk'}\alpha_{nk}\alpha_{nk'}q_{ik}q_{ik'}+\cdots+\lambda_{i,(k)}\prod_{k=1}^{K}\alpha_{nk}q_{ik}, \quad (2)$$

其中，$\lambda_{i,0}$是截距参数，它指未掌握项目 i 测量的任何属性而答对该项目的效应大小；$\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha}_n, \boldsymbol{q}_i)$ 表示的是属性主效应和交互效应。$\lambda_{i,k}$是在项目 i 上对应的属性 k 的主效应，掌握属性 k 对答对项目 i 的影响越大，则$\lambda_{i,k}$越大。$\lambda_{i,kk'}$表示的是同时掌握属性 k 和属性 k' 对正确作答项目 i 的效应大小。$\lambda_{i,(k)}$ 表示的是所有属性的交互效应。

2.2 Rasch 模型

Rasch 模型（Rasch, 1960）是当前最常用的项目反应模型之一，由丹麦学者 Rasch 提出，其表达式如公式（3）：

$$P(X_{ni}=1)=\frac{exp(\theta_n-b_i)}{1+exp(\theta_n-b_i)}, \quad (3)$$

其中，θ_n为被试 n 的潜在特质水平；b_i为项目 i 的难度参数。Rasch 模型的最大特点是它只有难度一个项目参数，具有参数少、较实用等优点而受到研究者们的广泛使用（罗照盛，2012）。

2.3 项目反应树模型

项目反应树模型（IRTree; Debeer et al., 2017）是具有树结构的项目反应模型，包括内部节点、分支和叶子三个部分。反应树中每个节点分支的概率，可以用 IRT 模型表示。而被试产生最终反应类别的概率即为通过的每个节点对应分支的概率乘积。IRTree 模型的优点在于能够结合个体的认知理论建模，可以在一定程度上体现被试产生最终作答结果可能经历的内部心理过程。反应树模型可以是线性的，也可以是嵌套的，如图 1 所示。

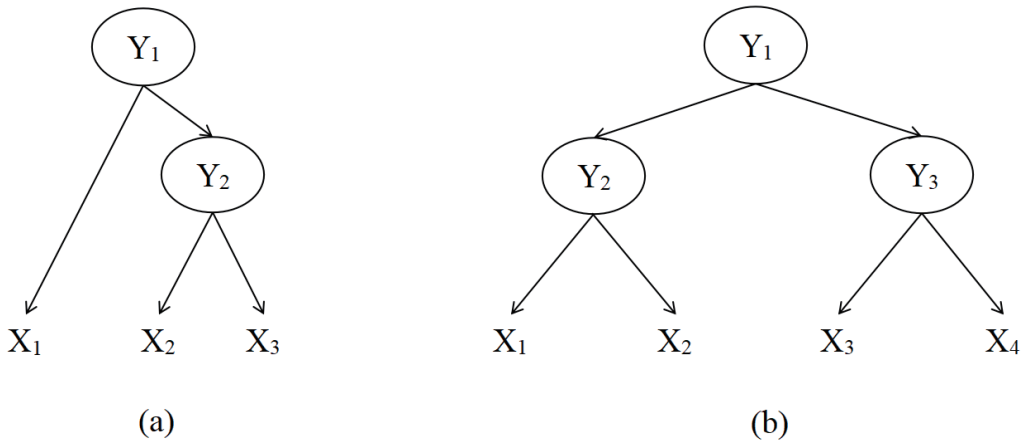

图 1 线性型 IRTree（a）和嵌套型 IRTree（b）

3 考虑缺失作答和随机猜测的认知诊断模型（IRTree-LCDM）开发

本研究借鉴项目反应理论领域中的项目反应树模型，用内部节点反映被试做出不同作答反应的过程，模型框架如图 2。模型假设被试在测验的项目上有两类表现，一类是正常作答，一类是异常作答。当被试决定异常作答时，极有可能做出两种反应：放弃作答此项目导致缺失或随机猜测此项目。参考前人研究，如 Debeer 等人（2017）、Huang（2020）和 Ulitzsch 等人（2020b）的模型，对是否缺失和是否猜对两个节点做出如图 2 的建模结构。

图 2 中，$P(X^{(R)})$为正常作答的概率，$P(X^{(A)})$为正确答对的概率，$P(X^{(M)})$为缺失作答的概率，$P(X^{(G)})$为猜对的概率。1 表示观察结果为答对，0 表示观察结果为答错，*MISS* 表示被试缺失作答。被试对当前项目进行作答涉及四个节点：是否正常作答、是否答对、是否缺失、是否猜对。表示被试在当前节点上的反应，所有节点都有 $X=0$ 和 $X=1$ 两个分支。

第一个节点$X^{(R)}$表示被试在当前项目上是否正常作答，使用 Rasch 模型来表达，将被试的作答反应分为正常作答和异常作答两类。当$X^{(R)}=1$，说明被试正常作答了该项目；反之，当$X^{(R)}=0$ 时，说明被试进行了异常作答。根据 Rasch 模型，被试在当前项目上正常作答的概率为：

$$P\left(X_{ni}^{(R)}=1\right)=\frac{exp\left(\theta_n^R-b_i^R\right)}{1+exp\left(\theta_n^R-b_i^R\right)}, \quad (4)$$

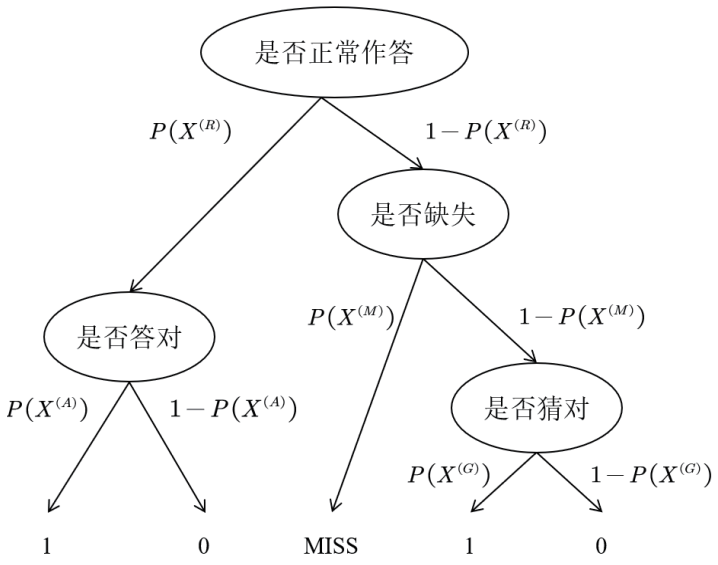

图 2 IRTree–LCDM 框架图

其中，b_i^R 表示题目 i 的正常作答阈值，θ_n^R 为被试 n 的正常作答倾向。

第二个节点为答对或答错节点$X^{(A)}$，使用饱和的认知诊断模型——LCDM 表达，即被试答对的概率函数采用 LCDM。具体公式如下：

$$P\left(X_{ni}^{(A)}=1|\boldsymbol{\alpha}_n\right)=\frac{exp\left(\lambda_{i0}-\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_n,\boldsymbol{q}_i)\right)}{1+exp\left(\lambda_{i0}-\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_n,\boldsymbol{q}_i)\right)},\quad(5)$$

其中，

$$\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_n,\boldsymbol{q}_i)=\sum_{k=1}^{K}\lambda_{ik}\alpha_{nk}q_{ik}+\sum_{k=1}^{K-1}\sum_{k=k+1}^{K}\lambda_{ikk'}\alpha_{nk}\alpha_{nk'}q_{ik}q_{ik'}+\ldots+\lambda_{i(k)}\prod_{k=1}^{K}\alpha_{nk}q_{ik},\quad(6)$$

其中，α_{nk}表示了被试 n 是否掌握了属性 k，如果被试掌握了属性 k，那么α_{nk}=1 ，否则 α_{nk}=0。K 为项目 i 测量的属性数量，λ_{i0}是截距参数，表示项目 i 测量的任意属性被试 n 都没有掌握，在此情况下依然答对该项目的效应大小； λ_{ik}是项目 i 所测量的属性 k 的主效应，表示被试 n 掌握该属性在答对该项目上的作用大小；$\lambda_{ikk'}$是指项目 i 上属性 k 和属性 k' 的交互效应，表示同时掌握这两个属性对答对该项目的贡献大小。$\lambda_{i(k)}$为掌握所有属性对答对项目的效应。被试答对项目i的概率为项目的截距效应、主效应和交互效应的和的 logit 形式。理论上来说，根据 LCDM 的性质，针对当前项目测量的属性，被试掌握的属性越多,那么答对该项目的概率就越大。

第三个节点缺失作答节点$X^{(M)}$表示被试作答是否缺失，它由被试的缺失作答倾向和项目的缺失阈值决定，采用 Rasch 模型进行建模。被试 n 在项目 i 上缺失作答的概率为：

$$P\left(X_{ni}^{(M)}=1\right)=\frac{exp\left(\theta_n^M-b_i^M\right)}{1+exp\left(\theta_n^M-b_i^M\right)},\quad(7)$$

其中，b_i^M表示项目 i 的缺失阈值，θ_n^M为被试 n 的缺失作答倾向。

第四个节点为随机猜测节点$X^{(G)}$。根据 LCDM 的理论解释，项目的截距参数为没有掌握任何属性的人答对此题的效应或概率。因此在本研究中，被试随机猜对的概率定义为没有掌握该项目所有属性而答对的概率（Hsu et al., 2020），即 LCDM 中截距参数的 logit 形式，也即：

$$P\left(X_i^{(G)}=1\right)=\frac{exp(\lambda_{i0})}{1+exp(\lambda_{i0})}.\quad(8)$$

在本模型框架下，被试的作答呈现出三种结果，正确作答项目、错误作答项目和缺失作答。设Y_{ni}为被试 n 在题目 i 上最终作答的结果，Y_{ni} = 1 表示被试正确作答项目；Y_{ni} = 0 表示被试错误作答项目；Y_{ni} = *MISS* 表示被试在项目上缺失作答，则相应的概率表达为：

（1）被试正确作答的概率

根据模型图 2 可以发现，被试正确作答项目的概率为在正常作答情况下答对的概率与在异常作答情况下随机猜对的概率的和，即：

$$P(Y_{ni}=1)=P\left(X_{ni}^{(R)}=1\right)P\left(X_{ni}^{(A)}=1\right)+\left[1-P\left(X_{ni}^{(R)}=1\right)\right]\left[1-P\left(X_{ni}^{(M)}=1\right)\right]P\left(X_i^{(G)}=1\right)\quad(9)$$

（2）被试错误作答的概率

被试错误作答项目的概率为在正常作答情况下答错的概率与在异常作答情况下随机猜错的概率相加，即：

$$P(Y_{ni}=0)=P\left(X_{ni}^{(R)}=1\right)\left[1-P\left(X_{ni}^{(A)}=1\right)\right]+\left[1-P\left(X_{ni}^{(R)}=1\right)\right]\left[1-P\left(X_{ni}^{(M)}=1\right)\right]\left[1-P\left(X_{i}^{(G)}=1\right)\right]. \quad (10)$$

（3）被试缺失作答的概率

被试缺失作答的概率为被试选择进行异常作答的概率与在异常作答情况下决定缺失的概率的乘积，即：

$$P(Y_{ni}=MISS)=\left[1-P\left(X_{ni}^{(R)}=1\right)\right]\times P\left(X_{ni}^{(M)}=1\right). \quad (11)$$

需要说明的是，当测验中既无缺失作答也无猜测作答，那么本研究新开发 IRTree-LCDM 就可简化为传统 LCDM，即传统 LCDM 是本研究开发的模型的一个特例。同时，本研究中采用 LCDM 作为认知诊断模型，也可以采用其他的认知诊断模型（如 G-DINA，ACDM，DINA 或 DINO 等）进行替换，因此新模型更为灵活、适用性更强。

4 IRTree-LCDM 的参数估计方法

本研究采用贝叶斯马尔科夫蒙特卡罗（Markov Chain Monte Carlo, MCMC）算法来实现新模型 IRTree-LCDM 的参数估计。

对模型进行参数估计首先需要确定参数的先验信息，参数的先验分布如下：

被试正常作答倾向与缺失作答倾向服从二元正态分布：

$$\begin{pmatrix}\theta_n^R\\ \theta_n^M\end{pmatrix}\sim\left(\begin{pmatrix}\mu_{\theta^R}\\ \mu_{\theta^M}\end{pmatrix},\sum_{person}\right), \quad (12)$$

其中，μ_{θ^R}与μ_{θ^M}是被试正常作答倾向和缺失作答倾向的均值，通常设为 0，$\sum_{person}$服从自由度为 2，精度为单位矩阵的逆 Wishart 分布（Huang, 2016, 2020），即：

$$\sum_{person}\sim InvWishart\left(L_{person},2\right), \quad (13)$$

其中L_{person}是一个 2×2 的单位矩阵。

项目正常作答阈值与缺失作答阈值使用弱先验的信息，采用 Ulitzsch 等人（2020a, 2020b, 2020c）的做法，即项目正常作答阈值服从均值为 R，方差为$\frac{1}{O_R}$的正态分布；项目的缺失作答阈值服从均值为 M，方差为$\frac{1}{O_M}$的正态分布，具体设置如下：

$$b_i^R\sim N\left(R,\frac{1}{O_R}\right), \quad (14)$$

$$b_i^M\sim N\left(M,\frac{1}{O_M}\right), \quad (15)$$

其中，超参数 R 和 M 都服从均值为 0，方差为 10 的弱先验正态分布。超参数 O_R 和 O_M 都服从形状参数为 0.01，尺度参数为 1000 的伽马分布，具体设置如下：

$$R\sim N(0,10), \quad (16)$$

$$M\sim N(0,10), \quad (17)$$

$$O_R\sim Gamma(0.01,1000), \quad (18)$$

$$O_M\sim Gamma(0.01,1000). \quad (19)$$

其余项目参数，即截距参数、主效应、交互效应服从一个多元正态分布，具体设置如下：

$$\begin{pmatrix}\lambda_{i0}\\ \lambda_{ik}\\ \lambda_{ikk'}\\ \cdots\\ \lambda_{i(K)}\end{pmatrix}\sim N\left(\begin{pmatrix}\mu_{\lambda_{i0}}\\ \mu_{\lambda_{ik}}\\ \mu_{\lambda_{ikk'}}\\ \cdots\\ \mu_{\lambda_{i(K)}}\end{pmatrix},\sum_{item}\right), \quad (20)$$

其中，超参数$\mu_{\lambda_{i0}}$服从均值为 -2，方差为 0.5 的正态分布；超参数$\mu_{\lambda_{ik}}$服从均值为 4，方差为 0.5 的正态分布；$\mu_{\lambda_{ikk'}}$超参数服从均值为 0，方差为 0.5 的正态分布。具体设置如下：

$$\mu_{\lambda_{i0}}\sim N(-2,0.5), \quad (21)$$

$$\mu_{\lambda_{ik}}\sim N(4,0.5), \quad (22)$$

$$\mu_{\lambda_{ikk'}}\sim N(0,0.5), \quad (23)$$

$$\sum_{item}\sim InvWishart(L_{item},l_{item}), \quad (24)$$

其中，L_{item}是项目的多元单位矩阵，l_{item}是项目的自由度参数。

属性掌握模式的先验设置与 Zhan 等人（2019）一致，具体设置如下：

$$c_n\sim categorical(\pi[1:C]), \quad (25)$$

$$\pi[1:C]\sim Dir(\Delta[1:C]), \quad (26)$$

其中，C为所有可能的属性掌握模式，一般有 2^K 个，$c\in\{1,\cdots,C\}$。π是一个非负概率的权重向量。Δ 为 dirichlet 分布的向量参数，一般设置为 (1,1,……,1)。

5 研究 1 IRTree–LCDM 的性能验证

5.1 研究设计

本研究采用 3×2×2×4 四因素实验设计，即缺失作答比例（2.5%、5%、10%）× 随机猜测比例（2.5%、5%）× 样本量（1000、1500）× 处理方法（IRTree-LCDM、LCDM-FCS、LCDM-CIM、LCDM-ZR）。题目数量固定为 30 题，属性数量固定为 5 个。本实验的 Q 矩阵设置在 de la Torre（2011）测验 Q 矩阵的基础上增加了一个 R 矩阵的试题，以保证参数估计的可识别性，具体见表 1。

值得说明的是，由于 IRTree-LCDM 能够直接应对包含缺失值的数据，无需进行额外处理，因此在本研究中处理方法 IRTree-LCDM 指的是用 IRTree-LCDM 直接进行参数估计。而传统的 LCDM 在设计时仅考虑了 0 和 1 计分的正常作答情况，不能直接用于分析包含的缺失数据，因此用传统 LCDM 分析数据时需要先对缺失数据进行填补。在以往研究中（Cai et al., 2023; Dai, 2021），研究者通常采用完全条件规范（fully conditional specification, FCS）、修正项目均值（corrected item mean, CIM）替换和零替换（zero replace, ZR）等方法进行缺失数据的填补。在本研究中，处理方法 LCDM-FCS、LCDM-CIM、LCDM-ZR 分别表示用 FCS、CIM、ZR 三种方法填补缺失数据后，使用传统 LCDM 进行参数估计。接下来，本文将对这三种常见的缺失数据填补方法进行相应的说明，以便更全面地理解它们在处理缺失数据中的应用：

（1）FCS（Cai et al., 2023）方法是对缺失值多重插补的一种方法，其基本思想是依据观察到的变量的条件分布，通过迭代的方式对缺失数据进行填补。FCS 通过为每个部分观察到的变量定义一个单变量填补模型来确定联合分布$p\left(Y^{obs}, Y^{mis}\middle|\vartheta\right)$。填补模型通常是基于缺失变量的性质（例如连续型、半连续型、类别型和计数型）选择的一般线性模型。从一些简单的填充方法开始，如均值填补或随机抽取，FCS 算法重复迭代所有缺失变量的填补过程。具体来说，在第 t 次迭代中对于缺失变量包括以下抽样：

$$\vartheta_j^t \sim f\left(\vartheta_j\right) f\left(Y_j^{obs}\middle|Y_{-j}^{t-1}, \vartheta_j\right), \quad (27)$$

$$Y_j^{mis(t)} \sim f\left(Y_j^{mis}\middle|Y_j^{obs}, Y_{-j}^{t}, \vartheta_j^t\right), \quad (28)$$

其中，ϑ_j^t是第 j 个变量在第 t 次迭代时对应的模型参数，$f\left(\vartheta_j\right)$是参数$\vartheta_j^t$的先验概率分布。$f\left(Y_j^{obs}\middle|Y_{-j}^{t-1}, \vartheta_j\right)$表示给定其他所有变量$Y_{-j}$（除了第 j 个变量之外的所有变量）在第 t -1 次迭代的状态，是参数ϑ_j和第 j 个变量的观测部分Y_j^{obs}的概率分布。$Y_j^{mis(t)}$是第 t 次迭代时为第 j 个变量的缺失部分生成的填补值。$f\left(Y_j^{mis}\middle|Y_j^{obs}, Y_{-j}^{t}, \vartheta_j^t\right)$表示在给定第 j 个变量的观测部分Y_j^{obs}、其他所有变量Y_{-j}在第 t 次迭代的状态，是在当前参数ϑ_j^t的情况下，第 j 个变量的缺失部分Y_j^{mis}的概率分布。

（2）CIM（Dai, 2021）方法是利用被试已经完成的其他相关题目的平均分作为基准，并根据这些题目与待填补题目的关联强度对这个平均分进行适当调整，从而得到一个更为合理的填补值，其计算公式如下：

$$CM_{ni} = \left[\frac{\bar{y}_{n.}}{\frac{1}{I.obs_n}\sum_n \bar{y}_{.i}}\right]\bar{y}_{.i}, \quad (29)$$

其中，CM_{ni}为被试 n 在项目 i 上缺失值的替换值。$\bar{y}_{n.}$为被试 n 在测验中有效作答得分的均值，$\bar{y}_{.i}$为所有被试在项目 i 上有效作答得分的均值，$I.obs_n$为被试有效作答项目数。

（3）ZR（宋枝璘等 , 2022）方法即将缺失值计为 0 分。这种处理方式在实际的教育考试中十分普遍，例如在高考、中考及各类学业成就考试或选拔性考试中，缺失作答通常被视为 0 分处理。

5.2 模拟过程

5.2.1 被试参数模拟

根据 Q 矩阵设置的 5 个属性，生成所有可能的属性掌握模式（knowledge states, KS），共有 2^K 种，也就是 32 种。被试的属性掌握模式 α 从这 32 种 KS 里随机抽取。被试正常作答倾向和缺失倾向从一个二元正态分布中生成，其中均值为 0，方差协方差矩阵中的对角线元素为 1，非对角线的元素为 -.5。

5.2.2 项目参数模拟

项目参数的模拟借鉴了同类研究（Debeer et al., 2017; Huang, 2020; Ulitzsch et al., 2020b）的做法，具体为：项目参数，即截距参数、主效应、交互效应从多元正态分布中产生，其中多元正态分布的均值向量为 -2、4、0，方差协方差矩阵为非对角元素为 0.5 的单位矩阵。不同异常作答比例下的正常作答阈

表 1 测验 Q 矩阵

项目	属性				
	A1	A2	A3	A4	A5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	1	0	0	0
7	1	0	1	0	0
8	1	0	0	1	0
9	1	0	0	0	1
10	0	1	1	0	0
11	0	1	0	1	0
12	0	1	0	0	1
13	0	0	1	1	0
14	0	0	1	0	1
15	0	0	0	1	1
16	1	1	1	0	0
17	1	1	0	1	0
18	1	1	0	0	1
19	1	0	1	1	0
20	1	0	0	1	1
21	1	0	1	0	1
22	0	1	1	1	0
23	0	1	1	0	1
24	0	1	0	1	1
25	0	0	1	1	1
26	1	0	0	0	0
27	0	1	0	0	0
28	0	0	1	0	0
29	0	0	0	1	0
30	0	0	0	0	1

表 2 不同条件下正常作答阈值和缺失作答阈值设定

缺失比例	随机猜测比例	正常作答阈值	缺失作答阈值
2.5%	2.5%	-3.55、-3.45、-3.35、-3.25、-3.15	0.4
	5%	-3.1、-3.0、-2.9、-2.8、-2.7	1.2
5%	2.5%	-3.1、-3.0、-2.9、-2.8、-2.7	-0.4
	5%	-2.8、-2.7、-2.6、-2.5、-2.4	0.4
10%	2.5%	-2.45、-2.35、-2.25、-2.15、-2.05	-1.3
	5%	-2.25、-2.15、-2.05、-1.95、-1.85	-0.4

表 3 在 1000 人条件下模拟的数据中不同异常反应比例（单位：%）

条件	异常组合 1		异常组合 2		异常组合 3		异常组合 4		异常组合 5		异常组合 6	
试次	缺失	随机	缺失	随机	缺失	随机	缺失	随机	缺失	随机	缺失	随机
1	2.447	2.363	2.640	4.803	5.063	2.603	5.080	5.013	10.507	2.643	9.540	5.110
2	2.593	2.653	2.733	5.137	5.183	2.643	4.790	4.703	10.120	2.233	9.747	5.150
3	2.353	2.333	2.717	4.880	5.293	2.637	4.990	5.097	10.313	2.330	9.197	4.990
4	2.783	2.677	2.647	4.937	5.383	2.540	5.210	5.240	10.373	2.573	10.397	5.050
5	2.610	2.623	2.887	4.937	5.093	2.480	5.210	4.800	9.813	2.470	9.517	5.027
6	2.560	2.477	2.350	4.903	5.550	2.390	5.137	5.100	10.280	2.673	10.147	5.360
7	2.820	2.623	2.607	5.380	4.997	2.510	5.333	5.210	10.457	2.363	9.450	5.083
8	2.753	2.640	2.663	5.203	5.047	2.500	5.127	4.903	9.993	2.580	9.990	5.237
9	2.503	2.257	2.610	5.033	4.903	2.477	4.950	5.410	9.990	2.480	9.513	5.200
10	2.463	2.510	2.683	5.237	4.740	2.577	5.330	4.887	10.597	2.510	9.763	5.343
11	2.737	2.547	2.680	5.087	5.167	2.520	5.430	5.217	10.710	2.480	10.177	5.200
12	2.413	2.470	2.747	5.120	4.910	2.530	4.887	4.890	11.157	2.537	9.863	5.380
13	2.553	2.500	2.723	5.077	5.447	2.620	5.280	5.377	10.637	2.570	9.897	5.250
14	2.353	2.493	2.480	4.843	4.713	2.570	5.350	5.203	10.530	2.377	10.003	5.173
15	2.373	2.683	2.437	4.837	4.853	2.503	5.030	4.937	11.200	2.550	10.230	5.030
16	2.490	2.337	2.487	4.877	4.973	2.443	5.167	4.993	9.953	2.437	10.210	5.510
17	2.747	2.363	2.733	5.263	5.173	2.540	5.240	4.920	9.930	2.343	9.603	5.367
18	2.480	2.497	2.610	4.893	5.353	2.460	4.790	4.900	10.587	2.620	9.857	5.083
19	2.747	2.517	2.640	4.963	5.067	2.600	4.907	4.983	10.140	2.553	9.910	5.410
20	2.660	2.460	2.703	5.297	4.783	2.617	5.143	5.267	10.297	2.477	9.963	5.440
21	2.567	2.753	2.497	5.020	5.210	2.357	5.383	5.093	10.487	2.493	9.423	5.350
22	2.693	2.600	2.703	4.903	5.137	2.523	5.037	5.133	10.720	2.483	9.387	4.963
23	2.667	2.657	2.663	5.107	5.253	2.423	4.803	4.953	10.277	2.457	9.763	5.053
24	2.477	2.457	2.680	4.957	4.930	2.683	5.723	4.957	11.063	2.577	9.207	5.023
25	2.490	2.513	2.737	5.240	5.100	2.550	5.123	4.943	10.357	2.377	9.977	5.303
26	2.460	2.333	2.550	5.153	4.913	2.543	4.677	4.927	10.600	2.650	9.550	5.310
27	2.673	2.570	2.583	4.903	4.930	2.447	5.120	5.207	10.280	2.380	9.660	4.960
28	2.497	2.500	2.747	5.113	4.760	2.393	5.100	5.343	10.197	2.547	9.973	5.123
29	2.300	2.430	2.447	5.100	4.913	2.450	5.220	5.107	10.790	2.533	9.573	5.157
30	2.717	2.647	2.790	4.887	5.147	2.447	5.133	5.010	10.143	2.437	9.293	4.823
Mean	2.566	2.516	2.639	5.036	5.066	2.519	5.123	5.057	10.417	2.491	9.759	5.182

注：异常组合 1 为 2.5% 缺失 2.5% 随机；异常组合 2 为 2.5% 缺失 5% 随机；异常组合 3 为 5% 缺失 2.5% 随机；异常组合 4 为 5% 缺失 5% 随机；异常组合 5 为 10% 缺失 2.5% 随机；异常组合 6 为 10% 缺失 5% 随机。

值设为距离为 .4 的 5 个的等距值，重复 6 次后随机分配给各个项目，缺失作答阈值则对所有项目一致。此设置能够保证每次模拟出的作答数据中缺失作答比例和随机猜测的比例相对稳定。正常作答阈值和缺失作答阈值的具体设置如表 2。

5.2.3 作答矩阵模拟

根据被试参数和项目参数，使用公式（9）、（10）、（11）分别计算出每个被试在每个项目上正确作答、

错误作答和缺失作答的概率，来确定每个被试在每个题目上的作答结果，最终生成被试的作答矩阵。

表 3 是在 1000 人条件下模拟的 30 次数据中缺失作答和随机猜测占数据总数的比例。1500 人模拟的情况与 1000 人基本相似，这里未列出以节省篇幅。

5.3 参数估计与评价标准

使用软件 R 语言（Version 4.2.1 64-bit; R Core Team, 2022）和软件 JAGS（Plummer, 2003）自编程序来实现。模型参数估计设置 2 条马尔科夫链，每条链 25000 次迭代，每条链的后一半迭代用于模型参数推断。

参数估计的收敛性：程序计算潜在尺度缩减因子$\widehat{R}$来评估每个参数的收敛性，设定$\widehat{R}$小于 1.2 的值表示收敛（Brooks & Gelman, 1998; de la Torre & Douglas, 2004）。

参数估计精度：项目参数和被试参数估计精度评价指标为 Bias 和平均绝对离差指标（Mean Absolute Deviation; ABS），公式为：

$$Bias(\hat{\pi}) = \frac{\sum_{d=1}^{D}\sum_{j=1}^{J}\left(\hat{\pi}_{dj} - \pi_{dj}\right)}{D \times J}, \quad (30)$$

$$ABS(\hat{\pi}) = \frac{\sum_{d=1}^{D}\sum_{j=1}^{J}\left|\hat{\pi}_{dj} - \pi_{dj}\right|}{D \times J}, \quad (31)$$

其中，D 表示实验重复次数，J 表示项目数或被试量，$\hat{\pi}_{dj}$和π_{dj}分别表示参数在第 d 次实验中的参数估计值和真值。

被试属性的估计精度评价指标采用模式正确判准率（pattern correct classfication rate, PCCR）和属性边际判准率（attribute correct classfication rate, ACCR），公式为：

$$PCCR = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N}\left(I(\hat{\boldsymbol{\alpha}}_n, \boldsymbol{\alpha}_n)\right)}{D \times N}, \quad (32)$$

其中，$\hat{\boldsymbol{\alpha}}_n$和$\boldsymbol{\alpha}_n$分别为被试 n 估计得到的属性掌握状态和真实的属性掌握状态。当$\hat{\boldsymbol{\alpha}}_n = \boldsymbol{\alpha}_n$时，说明被试 n 的属性掌握状态的估计结果与真实的属性掌握状态完全一致，此时 I 值为 1，否则 I 值为 0。D 为实验次数，N 为实验被试数。PCCR 的范围为 0 到 1，PCCR 越大说明模型对被试的模式判准率越高。

$$ACCR = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N}\left(I(\hat{\alpha}_{nk}, \alpha_{nk})\right)}{D \times N}, \quad (33)$$

其中，$\hat{\alpha}_{nk}$和α_{nk}分别为被试 n 估计得到的属性 k 的掌握状态和真实的属性 k 的掌握状态。当$\hat{\alpha}_{nk} = \alpha_{nk}$时，说明被试 n 对属性 k 的掌握状态的估计结果与真实的属性 k 的掌握状态一致，此时 I 值为 1，否则 I 值为 0。D 和 N 的含义同 PCCR。ACCR 的值越大说明对被试单个属性的估计越好。

5.4 研究结果

5.4.1 IRTree-LCDM 参数估计精度验证

IRTree-LCDM 模型在各种实验条件下的收敛率在 .812 到 .882 之间，平均为 .851，表明模型参数估计基本收敛。在 1500 人条件下，模型参数估计的收敛率平均值为 .854，相比 1000 人条件下（收敛率平均值为 .848）有所提高，这一结果反映了样本量的增加在一定程度上能够改善模型参数估计的稳定性。

表 4 和表 5 分别是新模型 IRTree-LCDM 在 1000 人、1500 人时各实验条件下参数的返真性结果。在各实验条件下，项目参数的 Bias 在 -.046 至 .102 之间，大部分接近于 0，除交互效应这一参数外，其他项目参数的 ABS 都较小，在 .070 到 .352 之间，说明新模型项目参数估计精度较理想。与其他项目参数相比，交互效应这一项目参数估计精度较差，这一结果与前人研究相一致，在潜在建模过程中，交互效应参数的估计精度往往低于截距参数和主效应参数（Jiang et al., 2016; Peng et al., 2022）。

根据表 4 和表 5，新模型中θ^R和θ^M两个被试参数的估计精度也在可接受范围内，Bias 在 -.011 到 .007 之间，都接近于 0，ABS 在 .447 到 .697 之间，说明新模型对被试参数的估计精度尚可。随着缺失比例的增大，项目参数b^M和被试参数θ^M的估计精度也有所提高，b^M的 ABS 从 .126 至 .194 下降到了 .075 至 .116，θ^M的 ABS 从 .662 至 .697 下降到了 .616 至 .672。

对比表 4 和表 5 可以发现，随着样本量的增加，大部分模型的项目参数的 Bias 越接近于 0，ABS 越小。以项目参数b^M为例，在 2.5% 缺失 2.5% 随机猜测这一实验条件下，样本量为 1000 人时，Bias 为 .097，ABS 为 .194，而当样本量为 1500 人时，Bias 和 ABS 分别下降到 .007 和 .130。这说明样本量的增加有助于项目参数估计精度的提高。另外，在 1000 人条件下，被试参数的 Bias 在 -.011 到 .007

表 4 IRTree–LCDM 参数的返真性（1000 人）

参数	2.5%缺失				5%缺失				10%缺失			
	2.5%随机猜测		5%随机猜测		2.5%随机猜测		5%随机猜测		2.5%随机猜测		5%随机猜测	
	Bias	ABS	Bias	ABS	Bias	ABS	Bias	ABS	Bias	ABS	Bias	ABS
λ_0	.004	.173	-.005	.183	-.002	.184	.026	.186	.010	.181	.018	.191
λ_k	.008	.313	.007	.352	.008	.320	-.009	.343	-.012	.335	-.039	.340
$\lambda_{kk}...\lambda_{i(K)}$	-.029	.760	-.041	.735	-.013	.734	.060	.723	.005	.769	-.046	.784
b^R	.056	.151	.025	.135	.037	.106	.030	.116	.002	.083	.006	.089
b^M	.097	.194	.045	.126	.102	.145	.057	.122	.003	.135	.003	.109
θ^R	.007	.608	-.009	.590	.003	.535	.003	.529	-.001	.447	.006	.457
θ^M	-.011	.697	.001	.663	-.005	.678	.005	.642	.003	.672	-.007	.616

表 5 IRTree–LCDM 参数的返真性（1500 人）

参数	2.5%缺失				5%缺失				10%缺失			
	2.5%随机猜测		5%随机猜测		2.5%随机猜测		5%随机猜测		2.5%随机猜测		5%随机猜测	
	Bias	ABS	Bias	ABS	Bias	ABS	Bias	ABS	Bias	ABS	Bias	ABS
λ_0	-.004	.148	.001	.145	-.007	.149	.003	.153	.024	.165	.005	.159
λ_k	-.003	.264	.022	.286	.003	.277	-.008	.301	-.024	.291	-.038	.303
$\lambda_{kk}...\lambda_{i(K)}$	-.019	.700	.029	.730	.037	.723	-.009	.732	.003	.763	.002	.776
b^R	-.002	.117	.047	.125	.018	.101	-.006	.092	.004	.070	.000	.074
b^M	.007	.130	.084	.129	.059	.142	.007	.093	.030	.116	-.001	.075
θ^R	-.005	.601	-.005	.583	-.003	.536	.000	.526	-.004	.451	-.007	.453
θ^M	.005	.684	.004	.662	-.008	.675	-.004	.635	.003	.666	-.001	.621

之间，ABS 在 .447 到 .697 之间；在 1500 人条件下，被试参数的 Bias 在 -.008 到 .005 之间，ABS 在 .451 到 .684 之间，这一结果说明，样本量大小对被试参数的估计精度影响极小。

5.4.2 IRTree-LCDM 与 LCDM 诊断正确率比较

图 3 和图 4 分别是在 1000 人、1500 人条件下传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）与 IRTree-LCDM 被试属性判准率的结果比较。图 5 和图 6 分别是在 1000 人、1500 人条件下传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）与 IRTree-LCDM 模式判准率的结果比较。

在所有实验条件下，相较于传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换），新模型总是表现最好。另外，在三种缺失值填补方法中，零替换后的数据的诊断正确率是表现最差的。具体而言，新模型在不同异常比例条件下的 ACCR 均值在 .946 到 .965 之间，PCCR 在 .783 到 .851 之间，并且 ACCR 和 PCCR 的标准差在 .008 到 .023 之间，说明模型对被试属性掌握情况的估计比较稳定。使用 FCS 方法替换缺失值后，传统 LCDM 的 ACCR 均值在 .934 到 .963 之间，PCCR 在 .745 到 .841 之间；使用 CIM 方法替换缺失值后，传统 LCDM 的 ACCR 均值在 .923 到 .962 之间，PCCR 在 0.711 到 .839 之间；使用 ZR 方法替换缺失值后，传统 LCDM 的 ACCR 均值在 .917 到 .958 之间，PCCR 在 .696 到 .825 之间。这些结果表明，使用不同的缺失值处理方法对被试属性的诊断正确率有不同程度的影响，而新模型在应对缺失作答和随机猜测方面优势明显。

通过观察图 3、图 4、图 5 和图 6 可以发现，随着异常作答比例的增加，新模型和传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）的属性判准率和模式判准率均受到影响，表现出一定程度的下降。以 1000 人条件为例，在 2.5% 缺失 2.5%

随机猜测条件下，新模型的 ACCR 平均值和 PCCR 分别为 .965 和 .851，LCDM-FCS 的 ACCR 平均值和 PCCR 分别为 .963 和 .841，LCDM-CIM 的 ACCR 平均值和 PCCR 分别为 .962 和 .839，LCDM-ZR 的 ACCR 平均值和 PCCR 分别为 .958 和 .825；而在 10% 缺失 5% 随机猜测条件下，新模型的 ACCR 平均值和 PCCR 分别为 .947 和 .787，LCDM-FCS 的 ACCR 平均值和 PCCR 分别为 .935 和 .748，LCDM-CIM 的 ACCR 平均值和 PCCR 分别为 .924 和 .713，LCDM-ZR 的 ACCR 平均值和 PCCR 分别为 .918 和 .701。相比传统 LCDM，新模型受异常作答比例的影响最小，对属性的估计准确性更加稳定。

通过图 3、图 4、图 5 和图 6 可以看出，当数据存在缺失作答和随机猜测两种异常作答反应时，不论是单个属性的判准率还是模式判准率，新模型均优于传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）。

6 研究 2 IRTree-LCDM 的应用研究

6.1 数据描述

本实证研究选取了 2019 年 TIMSS 八年级数学测验项目中的 16 个题目，测验被试为美国学生。这 16 个题目均为 0、1 计分题目，测量了 7 个属性，其中数字（number）、代数（algebra）、几何（geometry）、数据和概率（data and probability）为内容域（content domains）属性，理解（knowing）、应用（applying）、推理（reasoning）为认知域（cognitive domains）属性，题目 Q 矩阵如表 6 所示。由于 TIMSS 测验对作答反应的编码非常详细且多样化，涵盖了各种作答和非作答情形。为了使数据更契合研究二的分析需求，研究者需要对数据进行事先清理，剔除所有不符合分析要求的作答编码。例如，TIMSS 测验中编码 6 和编码 96 表示未到达（not reached）、系统缺失（sysmis）以及未施测（not administered）的作答，

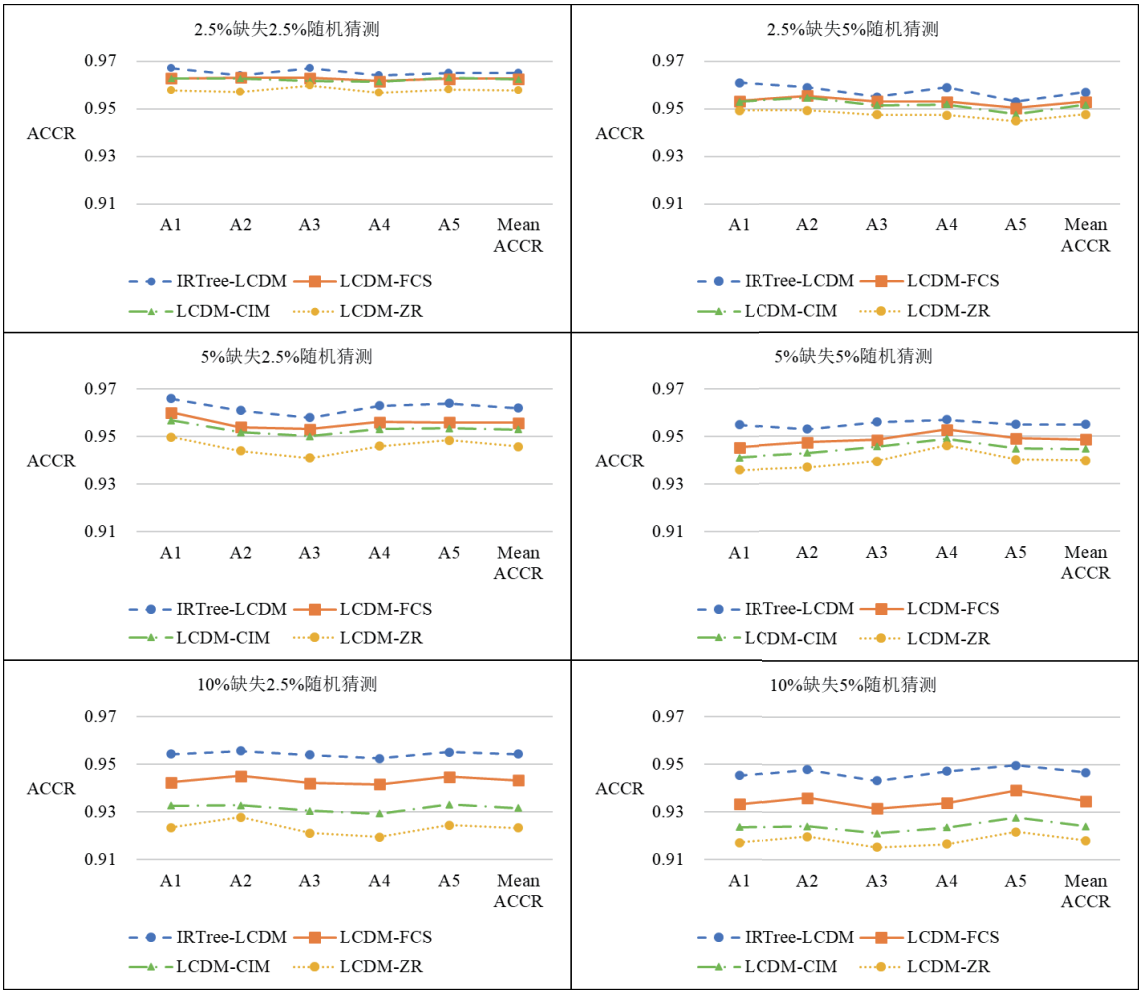

图 3 IRTree–LCDM 与传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）的 ACCR 比较（1000 人条件下）

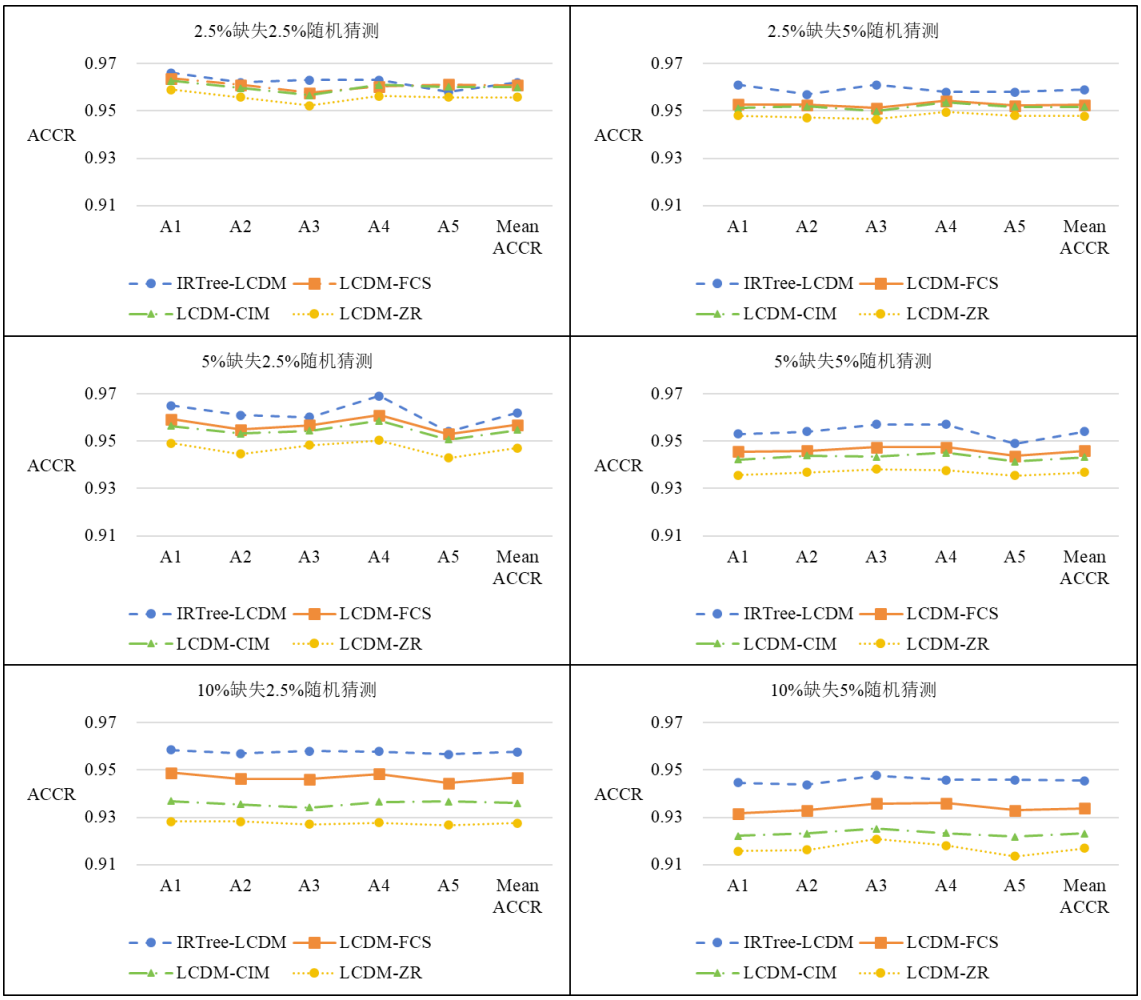

图 4 IRTree–LCDM 与传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）的 ACCR 比较（1500 人条件下）

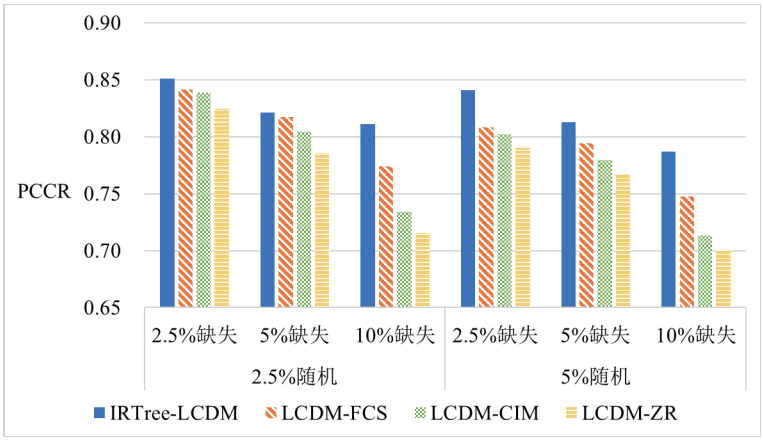

图 5 IRTree–LCDM 与传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）的 PCCR 比较（1000 人条件下）

这些作答反应不在本研究关注的异常作答类型范围内。因此，这些数据应被排除在外。经过数据清理后，最终保留的数据仅包含了 0 分（错误作答）、1 分（正确作答）和缺失作答三种反应形式。清理后的有效样本总数为 1221 人，其中缺失作答的比例为 1.51%。本研究中的数据和 Q 矩阵可以从 https://www.iea.nl/data-tools/repository/timss 下载。

6.2 研究方法

研究 2 选用传统 LCDM 与本研究提出的 IRTree-LCDM 进行数据分析并比较，参数估计方法与研究一相似。由于传统 LCDM 在设计时仅考虑了 0、1 计分的正常作答情况，因此对于缺失作答的数据

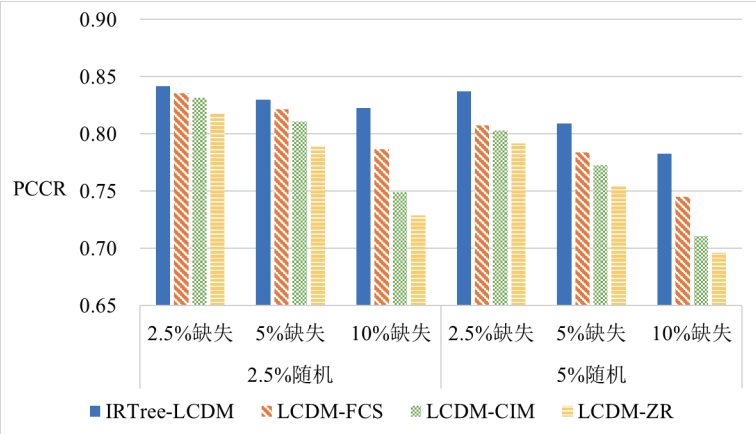

图 6 IRTree–LCDM 与传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）的 PCCR 比较（1500 人条件下）

表 6 TIMSS2019 中八年级数学测验 16 题的 Q 矩阵

题目	Number	Algebra	Geometry	Data and Probability	Knowing	Applying	Reasoning
ME52024	1	0	0	0	1	0	0
ME52058A	1	0	0	0	0	1	0
ME52058B	1	0	0	0	0	1	0
ME52125	1	0	0	0	0	0	1
ME52229	1	0	0	0	1	0	0
ME52063	0	1	0	0	0	1	0
ME52072	0	1	0	0	1	0	0
ME52146A	0	1	0	0	0	0	1
ME52146B	0	1	0	0	0	0	1
ME52092	0	1	0	0	0	1	0
ME52046	0	0	1	0	0	0	1
ME52083	0	0	1	0	0	1	0
ME52082	0	0	1	0	0	1	0
ME52161	0	0	0	1	0	1	0
ME52418A	0	0	0	1	0	1	0
ME52418B	0	0	0	1	0	1	0

需要进行处理，处理方法与研究 1 一致，即分别用 FCS、CIM 和 ZR 对缺失数据进行填补后用传统 LCDM 进行参数估计。IRTree-LCDM 则直接使用清理后的数据进行参数估计。

6.3 评价指标

本节研究使用 DIC（deviance information criterion）（Spiegelhalter et al., 2002）指标评价比较两个模型的拟合程度。DIC 是一个相对拟合指标，DIC 越小说明模型拟合更好。DIC 的计算公式如下：

$$DIC = \overline{D} + p_e, \quad (34)$$

其中，$\overline{D}$是指贝叶斯后验平均偏差，即偏差的后验期望，p_e是指模型中有效参数的数量。

6.4 研究结果

研究结果显示，IRTree-LCDM 中有 98.6% 的模型参数的$\hat{R}$值小于 1.2，表明参数估计已达到收敛。此外，IRTree-LCDM 的 DIC 值为 26436.80，LCDM-FCS 的 DIC 值为 31397.34，LCDM-CIM 的 DIC 值为 27915.27，LCDM-ZR 的 DIC 值为 26568.96，这一结果表明尽管使用一定的缺失值填补方法，传统 LCDM 仍然不能很好拟合真实数据，而 IRTree-LCDM 则表现出更好的适应性和效果。

为了进一步说明 IRTree-LCDM 在实证应用中的

优势，本研究选取了 6 个典型被试进行具体分析（见表 7）。需要说明的是，在三种缺失值填补方法中，传统 LCDM 模型对零替换后的数据的拟合结果最好，因此，选择使用零替换方法进行缺失值替换的传统 LCDM（LCDM-ZR）的估计结果与新模型的估计结果进行比较。具体而言，对于被试 A 和被试 B，他们的作答分数（15 分和 12 分）和正常作答倾向（$\theta^R = 4.51$ 以及 $\theta^R = 2.89$）均较高且没有缺失作答反应，这表明他们在大部分题目上都认真作答。在这种情况下，他们的作答数据完整，无需填补，IRTree-LCDM 和传统 LCDM 对被试的属性掌握模式的估计结果一致。被试 C 和被试 D 的作答分数中等，正常作答倾向较低，并且存在缺失作答（缺失量分别为 2 和 1）。由于传统的 LCDM 不具备处理缺失作答这一异常作答行为的能力，这导致它可能会错误地判定被试 C 没有掌握数字（number）属性，被试 D 没有掌握数据和概率（data and probability）属性，从而低估了被试的属性掌握数量。相反，IRTree-LCDM 能够更好地处理包含缺失的作答数据。如图 7 所示，被试 C 在 IRTree-LCDM 下掌握数字（number）属性的概率高于传统 LCDM，这表明 IRTree-LCDM 更准确地判定了被试 C 掌握了数字（number）这一属性。此外，被试 E 和被试 F 没有缺失作答，但是作答分数（均为 7 分）以及正常作答倾向（$\theta^R = .71$ 以及 $\theta^R = -.05$）均较低，这种情况表明这两个被试可能存在随机猜测作答，即被试在部分题目上作答不是根据自己能力水平做答而是猜测，由于 IRTree-LCDM 在构建时剔除了随机猜测对诊断的影响，因而判定被试掌握的属性数量比传统 LCDM 判定的更多。如图 7 所示，被试 E 掌握数据和概率（data and probability）属性的概率在 IRTree-LCDM 下比在传统 LCDM 下更高，因此判定被试 E 掌握了数据和概率（data and probability）两个属性。相比之下，IRTree-LCDM 对被试的属性掌握模式的估计更加合理，这进一步说明了 IRTree-LCDM 在应对缺失作答和随机猜测作答上具有一定的优势。

表 7 典型被试估计结果

被试	ID	得分	缺失量	IRTree-LCDM			LCDM-ZR 属性掌握模式
				正常作答倾向 θ^R	缺失作答倾向 θ^M	属性掌握模式	
A	50400906	15	0	4.51	-.66	1111111	1111111
B	52740608	12	0	2.89	-.18	1101110	1101110
C	52490110	10	2	1.15	2.62	1101110	0101110
D	50740610	11	1	1.67	1.60	1101110	1100110
E	52791003	7	0	.71	-.52	0101101	0100101
F	50211019	7	0	-.05	-.46	0111110	0011110

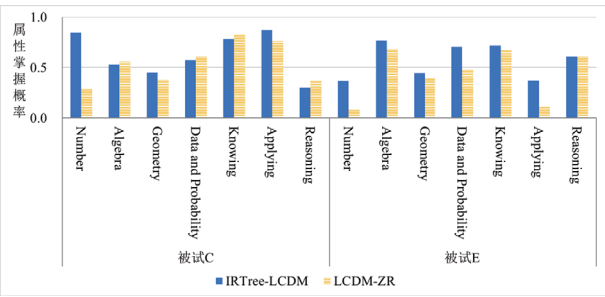

图 7 两个异常作答被试的属性掌握概率

7 总结与讨论

在心理与教育测验中，被试的异常作答反应，尤其是缺失和随机猜测作答，会导致一定的参数估计偏差，进而危害到测验的信度、效度乃至测验的公平性。但目前在认知诊断领域中针对异常作答反应的建模研究十分薄弱，这不利于认知诊断技术更好地服务实践。针对这一现状，受 IRTree 框架（Boeck & Partchev, 2012）的启发，本研究首次尝试将 IRTree 框架与认知诊断模型联合建模，开发了一个同时考虑缺失作答和随机猜测的认知诊断模型（IRTree-LCDM），其一是为了应对认知诊断中的异常作答，其二拟进一步提升传统认知诊断模型的诊断精度。

为了探讨并验证 IRTree-LCDM 的合理性、参数估计精度和实际应用价值，本文开展了两项研究：研究 1 采用了四变量（缺失作答比例、随机猜测比例、样本量以及处理方法）的模拟研究来验证 IRTree-LCDM 的参数估计精度以及 IRTree-LCDM 对被试属性和掌握模式的判准率，以深入探究样本量、缺失作答比例和随机猜测比例对模型参数估计效果的影响，从而反映 IRTree-LCDM 在应对缺失作答和随机猜测这两种异常作答的优势。研究 2 将 IRTree-LCDM 和传统的 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）应用于一个来自 2019 年 TIMSS 八年级数据测验的数据，通过对比模型数据的拟合效果并对典型被试进行分析，以体现 IRTree-LCDM 的实际应用价值。研究结果表明：（1）相较于传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）应对缺失数据的方法，本研究新开发的 IRTree-LCDM 具有较理想的参数估计精度及诊断正确率。（2）异常作答反应的比例会影响对属性及模式的判准率，异常作答反应的比例越高，判准率越低，但相比较传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换），新模型在应对缺失作答和随机猜测方面优势明显。（3）相较于传统 LCDM（使用零替换进行缺失值替换），IRTree-LCDM 在实际测验中的表现更好，对被试的属性掌握模式的估计更加合理。

虽然将 IRTree 框架与认知诊断模型联合建模取得了一定的成果，但本研究也存在一定的局限和不足，未来在以下几个方面还值得进一步研究和探索，以进步改善 IRTree-LCDM 并使其更好地服务实践：

（1）IRTree-LCDM 在处理缺失作答时主要考虑的是受到潜在缺失作答倾向影响的缺失反应。然而在实际测验中，尤其是在有时间限制的测验中，存在另一种不同的缺失情况，即被试由于时间不足而无法作答测验后期的题目，这种缺失是未经过认知加工的结果，与本文所考虑的缺失机制有所不同。因此，这种时间限制下的缺失作答需要与由于潜在的缺失作答倾向而导致缺失作答行为进行区分。未来的研究可以着重识别和区分这些不同的缺失机制，并采取适当的处理方法。通过更精确的建模和分析不同的缺失机制，可以进一步提高认知诊断评估的准确性，从而更好地反映被试的真实能力和作答状态。这将有助于推动认知诊断模型在实际应用中的有效性和可靠性。

（2）研究 1 的模拟实验所考虑的实验条件较为有限，因此未能对 IRTree-LCDM 进行全面的探究，具体表现在以下几个方面：（a）本研究考虑的两种异常作答比例有限，这虽然为模型评估提供了初步的视角，但这些比例可能不足以充分反映现实中可能遇到的情况。结果显示，随着异常比例的增加，IRTree-LCDM 和传统 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换）的属性判准率均有所降低，这一结果与前人研究一致（Peng et al., 2024），即当异常作答的比例提高时，模型的性能会受到影响，主要因为异常作答增加了噪声，减少了可用于参数估计的信息量，从而降低了估计的准确性。值得说明的是，尽管异常作答比例的提升会对两个模型的准确性均产生负面影响，但相较于传统的 LCDM（使用 FCS、CIM、ZR 三种方法进行缺失值替换），IRTree-LCDM 能够更有效地应对这种挑战，提供更准确的判准率。因此，未来的研究应重点探索如何在更高比例的异常作答情况下维持模型参数估计的准确性，并进一步提升模型在实际应用中的有效性和可靠性。（b）研究 1 的模拟实验仅考虑了属性之间的独立关系，然而在某些实际的测验中，属性之间可能存在层级关系。为了进一步评估 IRTree-LCDM 在属性存在层级关系时的表现，本研究新增了一个额外的实验条件，其中假设了 5 个属性之间存在分支型层级关系（见附图 1）。在这一条件下，IRTree-LCDM 的 ACCR 和 PCCR 分别达到了 .974 和 .886，明显优于传统 LCDM（使用零替换进行缺失值替换）的 .959 和 .838。此外，与属性独立的条件相比（IRTree-LCDM 的 ACCR 和 PCCR 分别为 .960 和 .833，使用零替换进行缺失值替换的传统 LCDM 的 ACCR 和 PCCR 分别为 .948 和 .796），属性间存在层级关系时两个模型的 ACCR 和 PCCR 均有所提高，这可能是因为层级关系简化了属性掌握模式的数量，从而提升了模型的判定准确性。未来的研究可以进一步探索不同类型的属性层级关系

对认知诊断测验中异常作答反应的影响，以便在更复杂的测验情境下提升模型的准确性和实用性。（c）研究 1 中考察的样本量均为较大样本，而在实际应用中，往往难以获得如此大规模的样本。对于 IRTree-LCDM 这样复杂的模型，其性能在小样本条件下可能会受到影响。本研究补充了 IRTree-LCDM 在样本量为 500 时的模拟实验，结果表明，尽管 IRTree-LCDM 的 ACCR 和 PCCR 仍然保持在较高水平（分别为 .958 和 .829），且优于使用零替换进行缺失值替换的传统 LCDM（ACCR 和 PCCR 分别为 .942 和 .773），但参数的收敛率显著下降，仅达到 78%。这一现象表明，在小样本情况下，IRTree-LCDM 可能难以稳定地进行参数估计，但 IRTree-LCDM 的参数估计仍有较高的精度（如项目参数的 Bais 平均值为 .091）。因此，未来的研究需要进一步探讨如何在小样本条件下提高 IRTree-LCDM 的参数收敛率，同时确保其参数估计的准确性。这可能涉及优化模型算法、增强数据质量或引入更为精细的统计方法，以增强模型在不同样本量下的稳健性和适应性。

（3）为了与前人的研究保持一致并减少模型的复杂度，新模型将猜测作答正确的概率约束为正常作答模型的截距参数。然而，实际上猜测行为可能受到多种因素的影响，因此这一假设可能过于简单。

（4）新模型仅考虑了 0/1 计分的项目，但在实际测验中，可能会包含多种类型的项目组合，例如多级计分题、多选题等。未来研究可以扩展模型，以适应更多类型的测验，扩大模型的适用范围。

（5）本文的模拟研究仅控制了样本人数和两类异常作答反应的比例，没有探讨属性数量和题目数量对模型参数估计精度的影响。因此，未来的研究应考虑这些因素对模型表现的潜在影响。

（6）随着测验电子化的发展，计算机化测验能够记录考生在作答时更多的过程性信息，这些信息在一定程度上可以辅助测验评估（von Davier et al., 2019）。未来可以将这些过程性数据纳入模型中，以更好地识别考生的异常作答行为，提高认知诊断模型的估计精度。

参考文献

刘玥 , 刘红云 . (2021). 心理与教育测验中异常作答处理的新技术 : 混合模型方法 . *心理科学进展 , 29*(9), 1696–1710.

罗照盛 . (2012). *项目反应理论基础* . 北京师范大学出版社 .

宋枝璘 , 郭磊 , 郑天鹏 . (2022). 认知诊断缺失数据处理方法的比较 : 零替换、多重插补与极大似然估计法 . *心理学报 , 54*(4), 426–440.

涂冬波 , 蔡艳 , 高旭亮 , 汪大勋 . (2019). *高级认知诊断* . 北京师范大学出版社 .

Boeck, P. D., & Partchev, I. (2012). IRTrees: Tree–based item response models of the GLMM family. *Journal of Statistical Software, Code Snippets, 48*(1), 1–28.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455.

Cai, M. Y., van Buuren, S., & Vink, G. (2023). Joint distribution properties of fully conditional specification under the normal linear model with normal inverse–gamma priors. *Scientific Reports, 13,* Article 644.

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

Dai, S. H. (2021). Handling missing responses in psychometrics: Methods and software. *Psych, 3*(4), 673–693.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Douglas, J. A. (2004). Higher–order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not–reached items using IRTrees. *Journal of Educational Measurement, 54*(3), 333–363.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*(6), 907–922.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika, 73*(1), 65–87.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log–linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Holman, R., & Glas, C. A. W. (2005). Modelling non–ignorable missing–data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*(1), 1–17.

Hsu, C. L., Jin, K. Y., & Chiu, M. M. (2020). Cognitive diagnostic models for random guessing behaviors. *Frontiers in Psychology, 11,* Article 570365.

Huang, H. Y. (2016). Mixture random–effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology, 7,* Article 1706.

Huang, H. Y. (2020). A mixture IRTree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement, 80*(6), 1168–1195.

Jiang, S. Y., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model.

Frontiers in Psychology, 7, Article 109.

Jin, K. Y., Siu, W. L., & Huang, X. T. (2022). Exploring the impact of random guessing in distractor analysis. *Journal of Educational Measurement, 59*(1), 43–61.

Kim, S., & Moses, T. (2018). The impact of aberrant responses and detection in forced-choice noncognitive assessment. *ETS Research Report Series, 2018*(1), 1–15.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement, 75*(5), 850–874.

Kuha, J., Katsikatsou, M., & Moustaki, I. (2018). Latent variable modelling with non-ignorable item non-response: Multigroup response propensity models for cross-national analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society, 181*(4), 1169–1192.

Liu, C. W. (2021). Examining nonnormal latent variable distributions for non-ignorable missing data. *Applied Psychological Measurement, 45*(3), 159–177.

Lu, J., Wang, C., Zhang, J. W., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology, 73*(2), 261–288.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*(2), 121–137.

Peng, S. W., Cai, Y., Wang, D. X., Luo, F., & Tu, D. B. (2022). A generalized diagnostic classification modeling framework integrating differential speediness: Advantages and illustrations in psychological and educational testing. *Multivariate Behavioral Research, 57*(6), 940–959.

Peng, S. W., Man, K. W., Veldkamp, B. P., Cai, Y., & Tu, D. B. (2024). A mixture model for random responding behavior in forced-choice noncognitive assessment: Implication and application in organizational research. *Organizational Research Methods, 27*(3), 414–442.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on Distributed Statistical Computing (DSC 2003)*(pp. 1–10).Vienna, Austria.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danmarks Paedagogiske Institut.

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Rios, J. A. (2022). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement, 82*(1), 122–150.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 64*(4), 583–639.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020a). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research, 55*(3), 425–453.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020b). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73(1)*, 83–112.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020c). A multiprocess item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement, 80*(3), 522–547.

von Davier, M., Khorramdel, L., He, Q. W., Shin, H. J., & Chen, H. W. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*(6), 671–705.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38.

Zhan, P. D., Jiao, H., Man, K. W., & Wang, L. J. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics, 44*(4), 473–503.

Cognitive Diagnostic Model for Miss and Random Guesses

Li Xiaopei[1], *Peng Siwei*[2], *Wang Qin*[1], *Cai Yan*[1]

([1] School of Psychology, Jiangxi Normal University, Nanchang, 330022)

([2] School of Psychology, Zhejiang Normal University, Jinhua, 321004)

Abstract With the advancement in psychological and educational testing, researchers have increasingly focused not only on measuring the abilities or traits of test takers, but also on assessing their mastery of specific knowledge structures. As a result, cognitive diagnostic assessment has become a major focus within the fields of psychological and educational measurement. In practice, however, both general and cognitive diagnostic tests frequently reveal abnormal response patterns from test takers, including missing responses and random guessing, which can be attributed to either individual characteristics or item properties. These abnormal responses can introduce biases in parameter estimation, thereby threatening the reliability and validity of the tests. Addressing these common abnormal response patterns is crucial for accurate data analysis. While much of the existing research on abnormal responses has been concentrated within the Item Response Theory (IRT) framework, there is a notable lack of work in the cognitive diagnosis domain, which remains in its early stages of development. Inspired by the IRTree framework, this study develops a novel cognitive diagnostic model that simultaneously accounts for missing responses and random guessing. This innovative model seeks to enhance the representation of abnormal response patterns within cognitive diagnostic assessments, offering significant implications for future research.

The paper begins with a comprehensive review of relevant concepts, theories, and prior research. It then details the modeling approach and framework of the new model, including the prior information for parameter settings and the Markov Chain Monte Carlo (MCMC) estimation method. A 3×2×2×4 four-factorial experimental design is employed, varying the proportions of missing responses (2.5%, 5%, 10%), proportions of random guessing (2.5%, 5%), sample sizes (1000, 1500), and handling methods (IRTree-LCDM, LCDM-FCS, LCDM-CIM, LCDM-ZR). This simulation study evaluates the parameter estimation accuracy and robustness of the new model and compares its attribute classification accuracy with traditional cognitive diagnostic models using different methods to handle missing values (i.e., full conditional specification, corrected item mean imputation, and zero replacement). Finally, the new model is applied to real data from the 8th-grade mathematics test of TIMSS 2019. The fit of the new model to the data is compared with that of traditional cognitive diagnostic models, and typical test-takers are analyzed to illustrate the advantages and practical value of the new model.

Results show that: (1)Compared to traditional LCDM using FCS, CIM, and ZR for handling missing data, the newly developed IRTree-LCDM exhibits superior parameter estimation and diagnostic precision. The average Attribute Classification Correct Rate (ACCR) for test takers exceeds 0.946, while the average Pattern Classification Correct Rate (PCCR) reaches .783. (2)The proportion of abnormal response patterns affects the classification accuracy of attributes and patterns; the higher the proportion of abnormal responses, the lower the classification accuracy. However, compared to traditional LCDM (using FCS, CIM, and ZR methods for missing data imputation), the new model shows significant advantages in handling missing responses and random guessing. (3)Compared to traditional LCDM (using ZR for missing data imputation), IRTree-LCDM performs better in actual tests, providing more reasonable estimates of test takers' attribute mastery patterns.

In conclusion, the IRTree-LCDM model demonstrates significant value and importance in handling abnormal responses.

Key words cognitive diagnosis, item response tree model, item response theory, miss, random guess