《心理学报》审稿意见与作者回应

题目：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

作者：宋枝璘，郭磊，郑天鹏

第一轮

审稿人 1 意见：

意见 1：在 CDM 的实践应用中，研究者经常会遇到数据缺失的问题，本研究比较了零替换、多重插补与极大似然估计法在 CDM 中处理缺失数据时的表现，具有一定的实用价值及创新性。

回应：感谢审稿人对本研究的肯定。

意见 2：根据作者文中陈述，MLE 包括 EM 及 FIML。“Dai(2018)在 DINA 模型基础上，仅比较了 EM 方法和其余传统方法的表现”“②在处理缺失数据时，未曾考虑到在缺失数据分析领域中表现更好的 MI 和 MLE 方法(Schafer & Graham, 2002)”这两句明显有冲突，请作者修改。

回应：感谢审稿人的意见，我们已将矛盾的表述加以改正，修改为：“但过往研究首先未曾考虑在缺失数据分析领域中表现较好、应用广泛的 MI 和 FIML 方法。”

意见 3：2.1 部分，存在符号使用问题，请修改。

回应：感谢审稿人的意见，我们已对 2.1 部分使用错误的符号进行了修改，如下所示：

“本研究所采用的诊断模型为 GDINA，其表达形式见公式(1)：

$$P\left(Y_{ij}=1\left|\boldsymbol{\alpha}_{ij}^{*}\right.\right)=\delta_{j0}+\sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{ik}+\sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1}\delta_{jkk'}\alpha_{ik}\alpha_{ik'}+\ldots+\delta_{j1,\ldots,K_j^*}\prod_{k=1}^{K_j^*}\alpha_{ik} \quad (1)$$

在 GDINA 模型中，被试在每道题目上被归为$2^{K_j^*}$个潜类别，其中$K_j^*=\sum_{k=1}^{K}q_{jk}$，表示题目 j 所考察的属性数量，$q_{jk}$=1 表示题目 j 考察了属性 k。$\boldsymbol{\alpha}_{ij}^{*}=\left(\alpha_{ij1},\ldots,\alpha_{ijK_j^*}\right)$为在被试属性向量

$\boldsymbol{\alpha_{ij}} = \left(\alpha_{ij1}, \ldots, \alpha_{ijK_j}\right)$基础上，仅保留题目 j 所考察属性，形成的坍塌(collapse)属性向量(K_j为测验考察的所有属性个数)。δ_{jo}为题目 j 的截距项，即当被试未掌握题目所考察属性时正确作答的基线参数。δ_{jk}为属性 k 的主效应，表示当被试仅掌握某一属性 k 时，对正确作答概率的影响。$\delta_{jkk'}$是题目 j 在属性 k 和 k'上的二阶交互效应，表示同时掌握两个属性对正确作答概率的影响。$\delta_{j1,\ldots,K_j^*}$为题目 j 在属性 1, 2,…, K_j^*上的最高阶交互作用，表示掌握了题目 j 考察的所有属性时，对正确作答概率的影响。其中，截距项δ_{jo}衡为非负数，主效应项为非负数，而交互作用项可以取任意值。”

意见 4：2.2 缺失数据的处理方法部分，各种缺失数据在 CDM 中具体是如何实现的，请更加具体陈述，尤其是 EM。

回应：感谢审稿人的意见。我们已经对缺失数据的处理方法进行了进一步的阐述和解释。对于其在 CDM 中的应用，简单来说就是 ZR 法，多重插补法和 EM 法都是通过各自的计算逻辑，将原有的包含缺失数据的数据集进行处理，填补成一个完整的数据集，再将这些处理后的完整数据集输入 GDINA 模型中，进行参数估计。而对于 FIML 方法，则是使用 R 中的 GDINA 包进行处理，FIML 方法为 GDINA 包的默认处理缺失值方法，即当输入的作答矩阵为包含缺失数据的矩阵时，FIML 方法并未填补缺失数据，而是在计算似然函数时将缺失位置剔除，只计算已作答数据的似然值。

此外，由于本研究主要着重于对 ZR、MI、EM 和 FIML 这几种发展较为成熟的缺失值处理方法进行比较，因此并未呈现它们的计算公式和示意图，这也并不是本研究的重点。修改稿中，我们着重阐述了这些方法的技术原理和实现步骤，同时，为了方便有需要的读者了解详细内容，我们将介绍各个方法的引文进行了标注。各方法的原理介绍请参见修改稿 2.3.3 部分。

意见 5：“3.1 研究设计”部分，对于饱和模型而言，200 及 400 人被试量偏少。

回应：回答：感谢审稿人的意见。根据目前 CDA 中已有研究所使用的样本量，可以将被试量分为小样本、中等样本以及大样本。其中，小样本一般在 100 人及以下(Chiu, et al., 2018; Ma & Jiang, 2021; Sessoms & Henson, 2018)；中等样本一般设置在 500 人左右(康春花等, 2015;

Chiu, et al., 2018; Pan & zhan, 2020)；大样本一般在 1000 人及以上(詹沛达等, 2015; De La Torre., 2011; Liu, et al., 2016)。本研究所使用的被试人数水平均属于中、大样本级别。同时，参考审稿人的意见，修改稿加入了 1000 被试的实验条件，以进一步对各缺失数据处理方法进行系统全面的比较。

1000 被试条件下的实验结果与 200、400 被试条件下的实验结果的总体趋势与变化规律基本相符，进一步支持了本文的实验结论。

下列文献为回复该问题所提到的相关文献：

康春花,任平 & 曾平飞.(2015).非参数认知诊断方法:多级评分的聚类分析.心理学报(08),1077-1088.

詹沛达,李晓敏,王文中,边玉芳 & 王立君.(2015).多维题组效应认知诊断模型.心理学报(05),689-701.

Chiu, C. Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. Psychometrika, 83(2), 355-375.

De La Torre, J. (2011). The generalized DINA model framework. Psychometrika, 76(2), 179-199.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M 2 statistic to evaluate the fit of cognitive diagnostic models. Journal of Educational and Behavioral Statistics, 41(1), 3-26.

Ma, W., & Jiang, Z. (2021). Estimating Cognitive Diagnosis Models in Small Samples: Bayes Modal Estimation and Monotonic Constraints. Applied Psychological Measurement, 45(2), 95-111.

Pan, Y., & Zhan, P. (2020). The Impact of Sample Attrition on Longitudinal Learning Diagnosis: A Prolog. Frontiers in psychology, 11, 1051.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. Measurement: Interdisciplinary Research and Perspectives, 16(1), 1-17.

意见 6：“（3）题目质量”部分，请具体给出各个符号的含义。

回应：回答：感谢审稿人的意见。我们已经对各个符号的含义进行了进一步的说明，修改为：“题目质量：参照 Ma 等人(2016)的设置包含三个水平：高质量、中等质量和低质量。题目为低质量时，参数设定为：$\mathrm{P}\left(Y_{ij}=1\left|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{0}\right.\right)\in U(0.05,0.15)$，$\mathrm{P}\left(Y_{ij}=1\left|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{1}\right.\right)\in U(0.85,0.95)$；题目为中等质量时，参数设定为：$\mathrm{P}\left(Y_{ij}=1\left|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{0}\right.\right)\in U(0.15,0.25)$，$\mathrm{P}\left(Y_{ij}=1\left|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{1}\right.\right)\in U(0.75,0.85)$；题目为高质量时，参数设定为：$P\left(Y_{ij}=1\left|\boldsymbol{\alpha}_{\boldsymbol{ij}}^{*}=\mathbf{0}\right.\right)\in U(0.25,0.35)$，$P\left(Y_{ij}=1\left|\boldsymbol{\alpha}_{\boldsymbol{ij}}^{*}=\mathbf{1}\right.\right)\in U(0.65,0.75)$。

$P(Y_{ij}=1|\boldsymbol{\alpha}_{ij}^{*}=\mathbf{0})$表示被试 i 未掌握题目 j 考察的所有属性时，答对题目的概率。$\mathrm{P}(Y_{ij}=1|\boldsymbol{\alpha}_{\mathbf{ij}}^{*}=\mathbf{1})$表示被试 i 掌握了题目 j 考察的所有属性时，答对题目的概率。其中，Y_{ij}为被试 i 在题目 j 上的作答情况，$\boldsymbol{\alpha}_{ij}^{*}$为在被试原属性向量基础上，仅保留题目 j 所考察属性形成的坍塌属性向量。”

意见 7：“3.2 模拟过程”部分，缺失数据处理中 MI 使用 MICE 包，其他方法是通过什么实现的，请具体说明。

回应：回答：感谢审稿人的意见。我们已经对各个缺失值处理方法的具体实现进行了说明，修改为：“使用 R 软件与 SPSS26.0 实现。首先，ZR 法通过自编 R 代码实现。MI 方法使用 R 软件中的 MICE 包(Buuren et al., 2010)完成。为了保证处理效果，参照 Chen 等(2020)的研究将 MI 插补次数设定为 20 次。其次，在使用 EM 方法插补缺失数据时，我们首先采用了 R 软件中 TestDataImputation 包中的 EMimpute 函数，但在数据规模大、缺失比例较高（例：1000 人、30 题、30%缺失率）时，R 软件无法运行。因此，最终选用了 SPSS26.0 进行 EM 插补处理。FIML 方法通过 R 软件中的 GDINA 包完成。”

意见 8：在实证研究中，项目参数标准误是如何计算的？是通过经验交叉相乘，观察信息矩阵还是三明治信息矩阵，请具体说明。

回应：感谢审稿人的意见。实证研究中项目参数标准误的计算使用 R 软件中 GDINA 包中的函数计算，其默认计算方式采用的是经验交叉相乘方法。此外，Philipp 等人（2018）的研究已证明，经验交叉相乘方法具有计算简单且稳健，在各种条件下表现均较好的优势。

下列文献为回复该问题所提到的相关文献：

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models.Journal of Educational and Behavioral Statistics,43(1),

意见 9：在 CDM 实践中，完全使用饱和模型分析数据是不恰当的，应该使用项目水平模型比较统计量进行，如 Wald 或 LR 等。

回应：感谢审稿人的意见。由于实证数据的 Q 矩阵较为特殊：每个题目仅考察了一个属性，故仅存在属性的主效应，属性间的交互作用并不存在。因此，无论选择简约模型还是饱和模型，补偿模型还是非补偿模型均一样。而为了同前文的模拟研究保持一致，我们在实证研究

中仍将 GDINA 包里的 model 设置为 GDINA 模型进行数据分析。

意见 10：实证数据中仅报告了近似拟合，其绝对拟合情况如何？

回应：感谢审稿人的意见，我们使用 GDINA 包对模型的绝对拟合指标进行了计算，补充了”M2”、”df”、”p”、”RMSEA2”及其”90%CI”几项指标，在实证研究部分添加了相关结果的报告。新添加指标的研究结果如下：

处理方法	绝对拟合指标				
	M_2	df	*p*	$RMSEA_2$	90%CI
ZR	16.69	12	**0.162**	0.023	[0,0.047]
MI-PMM	**13.81**	12	**0.313**	**0.014**	[0,0.042]
MI-LOGREG.BOOT	22.54	12	0.032	0.035	[0.01,0.056]
MI-CART	22.14	12	0.036	0.034	[0.009,0.056]
EM	17.19	12	**0.143**	0.024	[0,0.048]
FIML	22.64	12	0.031	0.035	[0.01,0.057]

……

审稿人 2 意见：

意见 1：论文系统的比较了在 CDM 的框架下几种缺失值填补方法的表现，作者同时使用了模拟数据和实证数据，使得文章结果较为可靠。论文将缺失值的填补方法用到了 GDINA 模型中具有一定的新意，另外，论文的结论也对后续的理论和实证研究有较好的指导意义。

回应：感谢审稿人对本研究的肯定。

意见 2：文章在前言的阐述过程中，有一个重要问题就是文章价值和意义的阐述，不够明确，换而言之，没有就要研究问题的重要性来做综述。作者尝试探讨在 CDM 中缺失值的处理方法问题，那么可能的创新点或价值就有集中，例如评审人列举三种。一是，前人有 CD 的缺失值处理方法，但不够系统不够全面，而本文就是要系统群面的比较，那前言就需要从前人不够系统全面的问题视角下入手。二是，前人没有在 GDINA 下全面探讨过，那研究综述就需要从前人探讨了哪些模型，这样的探讨为什么不够，本文从 GDINA 下为什么好来入手。

三是，CD 中没有探讨过部分方法，而本文探讨了，那综述部分就需要阐述，CD 中目前探讨过哪些。

回应：感谢审稿人的建议。修改稿中，我们综合参考了您给出的三条建议，以第一条为主线，将本研究的价值和意义做了重新梳理和重点阐述。

首先对于第一、三条建议，修改稿中的内容如下："尽管已有上述文献研究了 CDA 中的缺失数据问题，但过往研究首先未曾考虑在缺失数据分析领域中表现较好、应用广泛的 MI 和 FIML 方法。其中，MI 法已被证明其表现较为优异和稳健(Van Buuren, 2018; Schafer & Graham, 2002)，且于近年来被广泛用于缺失数据的处理中(Leacy et al., 2017; Rezvan et al., 2015)。FIML 采用"一步式"操作，直接使用带缺失值的作答数据进行模型拟合，比其它基于模型的方法更加便捷(Graham, 2009; Schafer & Graham, 2002)，此外，基于模型的方法表现更加出色，但在不同研究背景下的表现有较大差异，取决于具体的模型、数据和条件(Newman, 2003; Dai, 2018)。因此，有必要在 CDA 中系统地探索这些基于模型方法的表现，并与传统方法进行比较。"

其次关于第二条建议，文中有如下内容："基于系统全面比较缺失值处理方法这一主旨，本研究还做了如下推进：①Dai(2018)采用的 DINA 属于简约模型，它的非补偿模型特点往往与现实测验情景不符。而饱和模型，如 GDINA 模型(de la Torre, 2011)等受到了较多关注，并应用于多数研究中(Bai, 2020; 高旭亮等, 2018)，GDINA 不仅包含属性主效应，还将属性间交互作用考虑在内，更加符合现实情况，对实际测验拟合更佳。"

意见 3：在缺失值处理方法的介绍中，作者需要给出更多总结性的评论。用以说明为什么选择这几种方法，是使用较多，还是已经被证明较好，还是更适合数据。

回应：感谢审稿人的意见，我们在前言中对选择各方法的原因进行了介绍，具体文段如下："目前，缺失数据的处理方法主要包括两大类：一是传统处理方法，如具有代表性的零替换(Zero Replace, ZR)方法。ZR 法首先操作便捷，尤其在处理大规模数据时非常快速且在绝大多数统计软件上均可实现，其次不会造成被试的大量流失。因此，ZR 是研究者经常选用的方法之一，在 CDA 中也有使用（Aryadoust & Goh, 2001; Lee et al., 2011），且 ZR 方法被目前较多大型教育评估，如 PISA、TIMSS、PIRLS 所采纳(Xiao & Bulut, 2020)。虽然传统方法比较便捷，但会导致统计效力和参数估计精度的下降，因此有研究者并不建议使用(Dong & Peng, 2013; Enders, 2010)。第二类是基于模型的处理方法，近年来，随着统计技术不断发展，基于模型的处理方法相继被提出，并被证明处理效果要优于传统方法，因此越来越受到重视。

其中，极大似然估计(Maximum Likelihood Estimation，MLE)和 MI(Multiple Imputation, MI)方法的应用最广泛(Rotnitzky, 2008; Schafer & Graham, 2002)。MLE 是通过加工似然函数对缺失数据进行处理，包括期望最大化算法(Expectation-Maximization algorithm, EM)和全息极大似然估计方法(Full Information Maximum Likelihood, FIML)。对于 FIML、EM 和 MI 三种方法，均有研究证明其表现要优于传统方法(Graham, 2009; Jeličić et al., 2010;Van Buuren, 2018 ;Wothke, 2000)。”“其中，MI 法已被证明其表现较为优异和稳健(Van Buuren, 2018; Schafer & Graham, 2002)，且于近年来被广泛用于缺失数据的处理中(Leacy et al., 2017; Rezvan et al., 2015)。FIML 采用“一步式”操作，直接使用带缺失值的作答数据进行模型拟合，比其它基于模型的方法更加便捷(Graham, 2009; Schafer & Graham, 2002)，此外，基于模型的方法表现更加出色，但在不同研究背景下的表现有较大差异，取决于具体的模型、数据和条件(Newman, 2003; Dai, 2018)。因此，有必要在 CDA 中系统地探索这些基于模型方法的表现，并与传统方法进行比较。”

此外，我们还在“2.3 缺失数据的处理方法”部分添加了总结性的说明，以进一步解释我们选取的缺失值处理方法的理由，具体内容为：“依据前文综述，本研究选取了常见且被广泛使用的传统方法 ZR 法(Aryadoust & Goh, 2001; Lee et al., 2011)，以及基于模型处理中应用最为广泛(Rotnitzky, 2008; Schafer & Graham, 2002; Leacy et al., 2017; Rezvan et al., 2015)，处理缺失值效果更具优势(Rasmussen, 2007; Graham, 2009; Jeličić et al., 2010; Wothke, 2000)并且适用于二分变量插补(Marshall et al., 2010; Van Buuren, 2018, Xiao & Bulut, 2020)的 MI-PMM、MI-CART、MI-LOGREG.BOOT、EM 和 FIML 这几种方法。”

意见 4：更换图形示例点的颜色，无法清晰分辨其中的颜色和图形。

回应：感谢审稿人的意见。本文模拟研究涉及的自变量较多，且曲线重叠性较高，因而导致图像中曲线较难辩别。修改稿中已对图形进行了修改，具体修改为：①将图例及图像的颜色更换为对比性、分辨性更高的颜色；②将图例形状略作缩小避免遮挡图像。具体图像见文中 3.4 部分。

意见 5：需要增加关于数据缺失的三种机制的说明，论文中第一次出现已经是在研究设计中了。

回应：感谢审稿人的建议。我们已将三种缺失机制的说明添加在正文中第二章的 2.2 节。具体内容如下：

“2.2 缺失数据机制介绍

缺失数据可以通过缺失机制进行分类，Rubin(1976)定义了三种缺失的数据机制；完全随机缺失(missing completely at random, MCAR)，随机缺失(missing at random, MAR)和非随机缺失(missing not at random, MNAR)。在 MCAR 机制下，数据的缺失是完全随机的，不依赖于任何变量，即不论其它变量（如题目难度、区分度、被试能力值等）如何变化，数据产生缺失的概率都是均等的；在 MAR 机制下，数据缺失的概率并不是随机的，会受到数据集中已观测到的、不含缺失值的变量（如被试年龄、能力值等）的影响，但不受缺失数据自身的影响；在 MNAR 机制下，数据缺失的概率与缺失变量本身相关，如某一问题设计的过于敏感造成的缺失。

在心理教育测评中，这三种缺失数据的机制都有可能存在。Huisman 和 Molenaar(2001)认为，测评中缺失的作答数据是学生无意中报告的，因此将测评中的缺失数据视为 MCAR 机制下的缺失；还有研究者假设测评中存在 MAR 机制，因为数据的缺失与特定的个体特征有关(Lord, 1980; De Ayala el al., 2001; Finch, 2008) ;还有研究表明在某道题目上数据的缺失是受到题目本身特征的影响，即存在 MNAR 缺失机制(Brown et al., 2014)。”

意见 6：ZR 的填补效果不佳是可以预想到的，给缺失值替换上 0 相当完全没有借助其它信息，直接给了一个“错误”反应，换而言之，评审人认为这跟被试本来的知识状态水平有关，若缺失的被试都是知识状态掌握较差的被试，ZR 的填补可能会较好，反之则可能较差。

回应：感谢审稿人的专业意见。我们十分认可审稿人的观点，ZR 方法确实在处理缺失值中效果不佳，并且和被试本来的知识水平有关，文中我们在图 5 下方的“PCCR 的结果和讨论”部分进行了说明：“首先，综合题目参数和 PCCR 的结果，相较于 MAR 和 MCAR 机制，ZR 在 MNAR 机制下表现更好，这一现象的原因可能是：MNAR 机制下，缺失数据对应的原始作答为“0”（即答错）的可能性更高，即认为缺失的产生是由于被试无法作答，与被试的知识掌握状态有关；而 ZR 方法正好使用“0”替换缺失值，同样将缺失看作是由于被试不会作答产生的。因此，使用“0”替换缺失数据的 ZR 法正符合 MNAR 的缺失原理，ZR 法在 MNAR 机制下的表现更好。”十分感谢审稿人的宝贵意见。

意见 7：缺失值填补的基本逻辑是，缺失了就需要填上信息，然后怎么样填上更准确的信息。那另外一种更为简单粗暴的方式就是直接删掉，或者根本不理用缺失部分的信息。评审人则更感兴趣一个问题，填补后的效果，与完全不填补直接删除的效果哪个更好？

回应：感谢审稿人的专业意见。删除法主要包括成列删除（List Wise, LW）与成对删除（Pair Wise, PW）。其中，LW 指删除包含缺失值的完整个案，PW 指计算某一题目的相关参数时，删除在该题目上有缺失的所有被试的数据，保留当前题目所有可能获得的有效数据进行计算。

有研究者的结果显示：LW 和 PW 这两种方法通常会导致数据信息和统计能力的损失，是研究缺失数据的学者最不推荐的方法(Wilkinson, 1999)。首先，这两种方法处理缺失值时将会导致有偏差的模型估计(Dong & Peng, 2013; Peng et al., 2006)，最终导致统计能力的损失，处理效果不佳。其次，LW 和 PW 适用于被试量大且缺失率小的情况。对于缺失比例较大的情况，经 LW 处理后，被试量会大幅度减少，导致模型无法拟合。例如本研究中，在 1000 人、15 题、5 属性、30％缺失条件下，经 LW 处理后的作答矩阵仅剩下 20-30 人，无法进行参数估计。PW 则适用于计算题目均值等简单的计算，在处理 CDA 数据时，经 PW 删除后，由于删除了缺失的作答，可能出现不同被试的作答向量长度不同的情况，从而使作答矩阵不完整，无法输入模型中进行拟合。因此我们未考虑 LW 与 PW 方法。

综上所述，本研究并没有采用删除法进行缺失值处理。

下列文献为回复该问题所提到的相关文献：

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. SpringerPlus, 2(1), 1-17.

Peng, C.-Y. J., Harwell, M., Liou, S.-M., Ehman, L. H., & others. (2006). Advances in missing data methods and implications for educational research. Real Data Analysis, 31-78.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54(8), 594.

意见 8：实证研究部分中，因为没有题目和被试的真值，选用了模型比较的指标来说明填补方法的好坏，这可以作为一种方式，但总是觉得其支持力度不够。作者是否可以找到更多被试的数据，先用更多的数据做一个题目参数的估计，作为相对比较的基线。

回应：感谢审稿人的意见。针对实证研究而言，我们最初的设计是找一批被试人数较多的完整数据，然后人为进行缺失值的生成，用以比较不同缺失值处理方法的表现。这样做就可以把完整数据估计出来的结果当做“真值”来看。但后来否定了该做法，因为该做法并不是真正意义上的 real data 研究，因为缺失值是人为按照缺失比率生成的，严格意义上来说，仍属模拟研究范畴。因此，最终才确定采用目前的数据进行研究，并且该做法也有前人进行过类似的研究作为证据（Shan & Wang, 2020）。此外，为了进一步增加实证研究的效力，我们在报

告相对拟合指标的基础上，增加了绝对拟合指标作为实验的结果和依据。使用 GDINA 包对模型的绝对拟合指标进行了计算，并在实证研究部分添加了相关结果的报告。具体指标结果可以参照上文对第 1 位审稿人第 9 个问题的回答。

下列文献为回复该问题所提到的相关文献：

Shan, N., & Wang, X. (2020). Cognitive Diagnosis Modeling Incorporating Item-Level Missing Data Mechanism. Frontiers in psychology, 11, 3231.

意见 9：模拟的情境设置是否有前人研究的依据，如果有需要补充引用，如果没有可简要说明设置理由。

回应：感谢审稿人的意见，针对模拟情景设置的依据，我们已经在文章中添加标注了出来，分别为：

题目质量：参照 Ma 等人(2016)的设置包含三个水平：高质量、中等质量和低质量。

数据缺失机制：包括三种缺失机制： MCAR、 MAR 和 MNAR(De Ayala, 2001; Finch, 2008)；

数据缺失率：包括三个水平：10%、20%、30%(Dai, 2018)；

题目数量：包括两个水平：15 题和 30 题(Dai, 2018)；

被试数量：包括三个水平，200 人、400 人和 1000 人(Dai, 2018; de la Torre, 2011)。

第二轮

审稿人 1 意见：

意见 1：作者已经较好地回答了我的问题。我认为只有一个重要问题需要作者进一步探讨，根据先前研究（见下面参考文献），经验交叉相乘方法相对于错误的信息矩阵计算方法，优势明显。但是根据先前研究结论，在估计项目参数标准误时，观察信息矩阵及三明治信息矩阵可能是一个更优的选择。但是由于观察信息矩阵及三明治信息矩阵计算的复杂性，GDINA 软件包暂时没有包含在内，目前只有 dcminfo 软件包可以进行计算。

Liu, Y., Xin, T., & Jiang, Y. (2021). Structural Parameter Standard Error Estimation Method in Diagnostic Classification Models: Estimation and Application. Multivariate behavioral research. Liu, Y., Xin, T., Andersson, B.,& Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. British Journal of Mathematical and Statistical Psychology, 72, 18-37.

刘彦楼, 辛涛, 李令青, 田伟, 刘笑笑. (2016). 改进的认知诊断模型项目功能差异检验方法——基于观察信

息矩阵的 Wald 统计量. 心理学报,48, 588-598.

Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2019). Improved Wald Statistics for Item-Level Model Comparison in Diagnostic Classification Models. Applied Psychological Measurement, 43, 402-414.

回应：感谢审稿人的意见。根据以往的研究，不同的 SE 估计方法会产生不同的结果，相较于经验交叉相乘方法，三明治矩阵和观察信息矩阵可能是一个更优的选择。

但在使用 dcminfo 软件包估计标准误的过程中我们发现：1）dcminfo 包目前仅支持以 logit 形式连接的模型，因此非 logit 形式的参数需要进行转换后方可使用。2）FIML 方法并未对缺失位置的数据进行处理，而是在建立似然函数时忽略掉有缺失位置的数据，进而完成参数估计过程，但 dcminfo 包并不支持这种缺失值处理方法。基于上述问题，且实证研究关注的主要目的是通过参数估计，对几种不同的缺失处理方法进行比较，因此本研究中使用了 GDINA 包中的经验交叉相乘方法计算题目参数标准误，这也是研究者经常采用的计算方式。此外，受审稿人意见启发，我们认为开发出能够处理缺失值的信息矩阵是一个非常有意思的研究方向。

基于上述，我们在正文“4.2 评价指标”部分进行了讨论和说明，内容如下：“在估计 SE 时，采用不同的信息矩阵会得到不同精度的结果（Liu et al., 2021）。本研究采用 GDINA 包中的经验交叉相乘方法计算 SE，该方法的优点是操作便捷，且估计参数时表现较好，在 CDM 研究中也较常使用（De la Torre, J, 2009; Ná jera et al, 2021; Xu et al. 2020）。”。此外，我们在“5.3 研究局限及展望”部分进行了展望，内容如下：“此外，本研究使用了经验交叉相乘法计算实证数据的题目参数标准误。但一些研究指出在估计题目参数标准误时，观察信息矩阵及三明治信息矩阵也是常用且有效的方法（刘彦楼等, 2016; Liu et al., 2019）。因此，在后续研究中可以在缺失值领域进一步对比三种信息矩阵的表现，选取更适合的方法计算标准误。”

第三轮

主编意见：

意见 1：引言部分，作者在介绍 Dai (2018)研究内容的时候，提到 DINA 模型，但没有介绍 DINA 的全称和含义。请提供 DINA 更多的信息。

回应：感谢主编的意见。我们在正文中对 DINA 模型的含义、全称及参考文献等进行了解释与补充。首先，在引言部分首次提到 DINA 模型时，标注了其全称和对应文献如下：“(Deterministic Inputs, Noisy “and” Gate)”，“(Junker & Sijtsma, 2001)”，此外，我们检查了文

章中的名词，对 GDINA 模型的全称也进行了补充如下：“(Generalized Deterministic Inputs, Noisy “and” Gate)”。并且在引言中补充了内容：“对 GDINA 及 DINA 模型的介绍及含义参见 2.1 部分。”

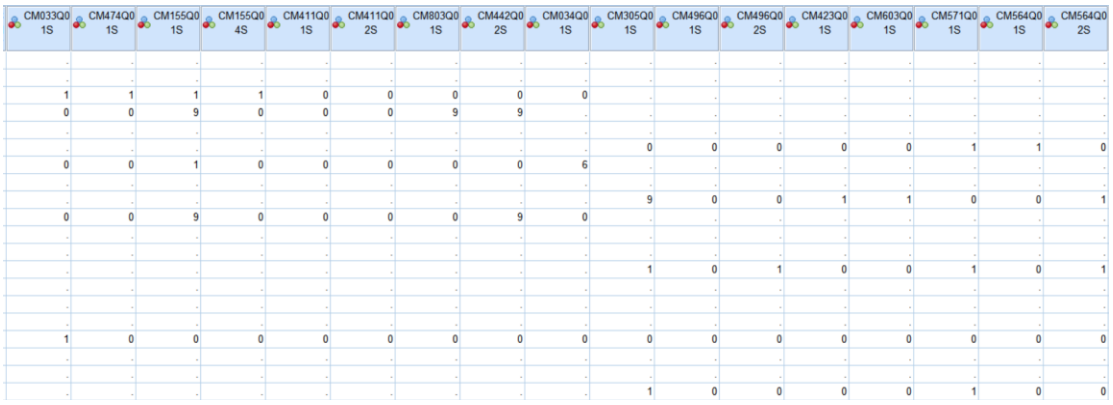
其次，在 2.1 认知诊断模型部分，我们对 DINA 模型的含义进行了概述如下：“GDINA 模型属于饱和模型，对 GDINA 进行约束，即仅保留公式（1）中的截距项和最高阶交互项，便可得到 DINA 模型：$P(Y_{ij}=1|\boldsymbol{\alpha}_{ij}^{*})=\delta_{j0}+\delta_{j1,\ldots,K_j^{*}}\prod_{k=1}^{K_j^{*}}\alpha_{ik}$。其含义为：当且仅当被试 i 掌握了题目 j 考核的所有属性时，该被试倾向于答对这道题目；而当被试 i 未掌握题目 j 考核的所有属性时，即认为该被试倾向于答错这道题目。”

意见 2：在 4.1 研究数据的部分，作者选择 9 道 PISA2015 的数学题目，但未阐明选择这 9 道题的原因。是否因为它们的缺失值更多，或者缺失值包含了 MCAR、MAR 和 MNAR 三种不用的缺失值种类?在模拟实验的条件中，作者设置了 15 和 30 两个题目数量水平，但应用到 PISA 数据时，作者只呈现了 9 道题，请阐明原因。此外，数据来源是多米尼加共和国，请阐明选择这一样本的原因。

回应：感谢主编的意见。首先，本研究选择 PISA2015 这批数据是由于：① 缺失比例合适。该批数据的缺失比例适中，因此能够展现出不同缺失值处理方法之间的差异。若缺失比例过小，使用不同缺失值处理方法得到的效果可能差异不会很明显，不方便比较与分析。② 需要已有文献或专家界定好实证数据的 Q 矩阵，以保证实证研究结果的有效性。③ PISA 属于国际大型测验，能够保证数据来源的合理性，这也是认知诊断研究中常选用其作为实证数据的原因(Chen & Chen, 2016; Wu et al., 2020; Zhan et al, 2018a; Zhou et al., 2021)。

其次，PISA2015 关于数学素养的认知诊断数据中，共有 17 道二级计分题目，其题目序号分别为 CM033Q01、CM474Q01、CM155Q01、CM155Q04、CM411Q01、CM411Q02、CM803Q01、CM442Q02、CM034Q01、CM305Q01、CM496Q01、CM496Q02、CM423Q01、CM603Q01、CM571Q01、CM564Q01 和 CM564Q02(Zhan et al, 2018b)。但在 PISA2015 实际测试时，不同被试所做题目组块不同，有些被试做前 9 题(CM033Q01、CM474Q01、CM155Q01、CM155Q04、CM411Q01、CM411Q02、CM803Q01、CM442Q02、CM034Q01)，有些被试做后 8 题(CM305Q01、CM496Q01、CM496Q02、CM423Q01、CM603Q01、CM571Q01、CM564Q01、CM564Q02)。因此，数据中会出现如下图所示的情况：同一位被试被分配到了前一组的九道题目，却没有被分配到后一组的八道题目，这会导致被试在某一组的所有题目

上出现缺失。



CM033Q01S	CM474Q01S	CM155Q01S	CM155Q04S	CM411Q01S	CM411Q02S	CM803Q01S	CM442Q02S	CM034Q01S	CM305Q01S	CM496Q01S	CM496Q02S	CM423Q01S	CM603Q01S	CM571Q01S	CM564Q01S	CM564Q02S
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	1	1	1	0	0	0	0	0	.	.	.	.	.	.	.	.
0	0	9	0	0	0	9	9	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	6	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	9	0	0	1	1	0	0	1
0	0	9	0	0	0	0	9	0	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	1	0	1	0	0	1	0	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	1	0	0	0	0	1	0	0

因此，如果将 17 道题目全部纳入分析，会出现问题：①被试要么整体缺失前 9 题作答，要么整体缺失后 8 题作答，均不符合本研究所探讨的三种缺失机制，因此，不同缺失值处理方法的比较便没有意义。②若将全部 17 道题目都进行分析，会导致总缺失比例过高（数据的总缺失率达 37.38%），原因正是由于有些被试并未作答前 9 题或后 8 题。根据本文模拟研究结果表明：缺失率较大时（如 30%），所有的缺失值处理方法均表现较差，此时的比较也没有任何意义。此外，本研究的模拟研究条件设置、以及实证数据分析均是参考了 Shan 和 Wang(2020)的研究进行设置，在他们的研究中也是选择了这 9 道题目作为实证数据。

之所以选择多米尼加共和国，原因在于：在参加了 PISA2015 的这 9 道题目的国家中，缺失率最小的是中国香港，缺失率为 1.40%；缺失率最大的是巴西，缺失率为 16.04%。较小的缺失率不利于展现不同缺失值处理方法的差异，也与本研究的模拟条件不符。因此，在比较多个国家的作答数据后，我们选择了缺失比率较为合适的多米尼加共和国（缺失率为 14.02%）的数据。

在 4.1 研究数据部分，我们进行了如下补充：“为进一步探讨不同缺失值处理方法的生态效度，本研究参考 Shan 和 Wang(2020)的实证研究，使用了 PISA2015 年基于计算机测评的数学测验数据作为实证数据，主要原因为：①缺失比例合适，能够展现出不同缺失值处理方法之间的差异。若缺失率较小，不同缺失值处理方法得到的效果可能差异不会很明显；缺失率较大时（如 30%），所有的缺失值处理方法均表现较差，此时的比较也没有任何意义。②具备已标定好的 Q 矩阵。③属于大型测验，结果可靠。”

下列文献为回复该问题所提到的相关文献：

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. Language Assessment Quarterly, 13(3), 218-230.

Shan, N., & Wang, X. (2020). Cognitive Diagnosis Modeling Incorporating Item-Level Missing Data Mechanism. Frontiers in psychology, 11, 3231.

Wu, X., Wu, R., Chang, H. H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. Frontiers in psychology, 11, 2230.

Zhan, P., Jiao, H., & Liao, D. (2018a). Cognitive diagnosis modelling incorporating item response times. British Journal of Mathematical and Statistical Psychology, 71(2), 262-286.

Zhan, P., Liao, M., & Bian, Y. (2018b). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. Frontiers in psychology, 9, 607.

Zhou, Y., Liu, Q., Wu, J., Wang, F., Huang, Z., Tong, W., & Ma, J. (2021). Modeling Context-aware Features for Cognitive Diagnosis in Student Learning. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining , 2420-2428.