《心理学报》审稿意见与作者回应

题目：CD-CAT 中基于 SCAD 惩罚和 EM 视角的在线标定方法开发——基于 G-DINA 模型
作者：谭青蓉; 蔡艳; 汪大勋; 罗芬; 涂冬波

第一轮

审稿人 1 意见：

本文基于正则化方法选择特征的思路提出一种新的 CD-CAT 在线标定方法（SCADOCM），新方法适用于 G-DINA 等广义认知诊断模型下的新题 Q 矩阵和题目参数同时性在线标定。另外，本文分别基于模拟题库和真实题库开展蒙特卡洛模拟研究，并从标定效率和标定精度两方面对 SCADOCM 和传统 SIE 方法进行比较。总的来说，研究有一定新意和较好的实践价值，而且内容完整、结构清晰。但是，论文在“公式符号”、“模拟细节”、“结果呈现和解释”和“文字表述”的准确性/完整性/规范性等方面还需要改进和完善。

意见 1：第 5 页提到“混合认知诊断模型”，什么是混合的认知诊断模型？请进行简单解释/说明。

回应：非常感谢专家的宝贵建议！这里将“饱和认知诊断模型”误写为“混合认知诊断模型”，在文章中已对此进行修改，并对饱和认知诊断模型进行了简要说明。详见文章第 4 页。

意见 2：第 6 页：（1）在对 alpha_c 中元素进行说明时，使用 alpha_ck 而不是 alpha_cK 进行说明；“一致性链接函数”在第一次出现时，就要给出其对应的英文“identity link function”（注意不是 identify link function）。

回应：非常感谢您的宝贵意见，我们在文章中已对上述问题进行修改。详见文章第 5 页。

意见 3：第 7 页：（1）“2.1 G-DINA 模型”最后一行对 delta_j 的公式描述不准确，包括的元素应该是 delta_j0, delta_j1, …, delta_j(Kj*), delta_j12, …, delta_j(Kj*-1)(Kj*), delta_j12…Kj*；（2）“…主要 JEA”表述不清楚。

回应：非常感谢您的意见。（1）我们已将 $\boldsymbol{\delta}_j$ 的公式描述修改为 $\boldsymbol{\delta}_j = (\delta_{j0}, \delta_{j1}, \cdots, \delta_{jK_j^*}, \delta_{j12}, \cdots, \delta_{j(K_j^*-1)(K_j^*)}, \cdots, \delta_{j12\cdots K_j^*})$。（2）将“…主要 JEA”修改为“CD-CAT 中已有的同时标定新题 *Q* 矩阵与项目参数的方法主要包含了 JEA (陈平, 辛涛, 2011b)、SIE (Chen et al., 2015)、IGEOCM (谭青蓉 等, 2021)和基于基尼的方法(Tan et al., 2022)等。”。详见文章第 6 页。

意见 4：第 8 页公式（2）中，如果要表示属性掌握模式为 alpha_c 的被试在新题 j 上的正确作答概率，应使用 P(q_j, delta_j; alpha_c) 或 P(q_j, delta_j | alpha_c)，P(q_j, delta_j, alpha_c) 一般表示三者的联合概率。下同。

回应：非常感谢您的宝贵意见。我们已将 $P(\boldsymbol{q}_j, \boldsymbol{\delta}_j, \boldsymbol{\alpha}_c)$ 统一修改为“$P(\boldsymbol{q}_j, \boldsymbol{\delta}_j|\boldsymbol{\alpha}_c)$”。

意见 5：第 9 页第 7~9 行，“…对于根据某一 q 向量估计的项目参数估计值，…，然后基于此标定新题的 q 向量与项目参数”这句话没有看明白，请给予更多的说明或解释。

回应：非常感谢您的意见！我们已对这句话进行了修改，并增加了例子来使表述更清晰，有助于理解。详见文章第 8 页。

意见 6：第 11 页公式（8）下一行的描述中，D 应该是大小为 K*nj 的属性边际掌握概率矩阵，否则无法进行矩阵运算。

回应：非常感谢您提出的宝贵意见！我们已将“***D*** 表示大小为$n_j \times K$被试属性边际掌握概率矩阵”修改为“***D*** 表示大小为$K \times n_j$的被试属性边际掌握概率矩阵”。

意见 7：第 13 页对项目质量的描述不准确。什么参数来自 U(0.05, 0.15)？什么参数来自 U(0.1, 0.3)？只呈现两个分布不完整。

回应：非常感谢您提出的宝贵意见！项目质量(高质量：$P_j(\mathbf{0})$ (未掌握项目 j 所测量的任一属性的被试在项目 j 上的答对概率)和 1- $P_j(\mathbf{1})$ (掌握项目 j 所测量的所有属性的被试在项目 j 上的答对概率)从$U(0.05, 0.15)$中随机抽取；低质量：$P_j(\mathbf{0})$和 1-$P_j(\mathbf{1})$从$U(0.1, 0.3)$中随机抽取)。我们已在文章 12 页补充该内容，使表述更完整。

意见 8：第 14 页对产生被试属性掌握模式的描述不完整。比如，从高阶分布和多元正态分布中抽出来的是连续值，如何离散化得到被试的属性掌握模式？

回应：非常感谢专家的宝贵意见！我们已在 13 页对产生被试属性掌握模式的描述进行修改，高阶分布和多元正态分布下属性掌握模式的生成分别如下：

高阶分布：“在高阶分布中，被试 i 是否掌握第 k 个属性与被试 i 的一般潜在能力 θ_i 有关，能力为 θ_i 的被试 i 掌握第 k 个属性的概率为

$$P(\alpha_{ik}=1|\theta_i,\lambda_{0k},\lambda_{1k})=\frac{exp(\lambda_{1k}\theta_i+\lambda_{0k})}{1+exp(\lambda_{1k}\theta_i+\lambda_{0k})}, \quad (15)$$

其中，λ_{0k} 和 λ_{1k} 为结构参数，研究中设置 K = 5，$\boldsymbol{\lambda}_0$ = (–1, –0.5, 0, 0.5, 1)，且对所有属性 k 均有 λ_{1k} = 1.5，被试 i 的能力值从 N (0, 1)中产生(de la Torre & Chiu, 2016)。在 0~1 之间生成一个随机数，将基于上式(公式 15)计算的概率值与随机数进行比较，若概率值大于随机数，被试 i 掌握属性 k，$\alpha_{ik}=1$，否则被试 i 未掌握属性 k，$\alpha_{ik}=0$ (Ma & de la Torre, 2020)。”。

多元正态分布：“在多元正态分布中，属性间的相关设置为 0.5 (J. Chen, 2017; Chiu, 2013)。假设被试 i 的能力向量为$\boldsymbol{\vartheta}_i=(\vartheta_{i1},\cdots,\vartheta_{iK})$，则被试 i 的属性掌握模式$\boldsymbol{\alpha}_i=(\alpha_{i1},\cdots,\alpha_{iK})$可通过以下公式获得 (Chiu, 2013)：

$$\alpha_{ik}=\begin{cases}1 & \text{if } \vartheta_{ik}\geq\Phi^{-1}\left(\dfrac{k}{K+1}\right)\\ 0 & \text{otherwise}\end{cases}, \quad (16)$$

其中Φ^{-1}是正态分布概率密度的逆函数。”。

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological*

Measurement, *41*(4), 277-293.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618.

Ma W., & de la Torre, J. (2020). "GDINA: An R Package for Cognitive Diagnosis Modeling." *Journal of Statistical Software*, *93*(14), 1–26.

意见 9：第 15 页对"CD-CAT 的模拟过程"可以改进。因为使用香农熵策略选题时不依赖于被试当前的知识状态估计值，所以被试每作答完一个题目后可以不用 MLE 估计被试的知识状态，而只需在测验结束后使用 MLE 方法估计一次被试的知识状态。

回应：非常感谢您的建议，我们已对 CD-CAT 的模拟过程进行修改。具体如下：

"(1)从题库中随机挑选一个项目作为被试的初始作答题; (2)模拟被试在当前项目上的作答, 然后基于被试在已选项目上的作答使用香农熵(shannon entropy, SHE; Cheng, 2009)选题策略为被试从剩余题库中挑选最适合的项目作为其下一个作答项目，重复该步骤直到测验长度达到预先指定的标准。SHE 选题策略理论基础扎实，具有较高的估计精度，已有同时标定新题 *Q* 矩阵和项目参数的研究也表明 SHE 选题策略下各在线标定方法均具有较好的项目标定精度(Chen et al., 2015; Tan et al., 2022; Zheng & Chang, 2016; 谭青蓉 等, 2021; 张学工, 2010)。因此，研究选用 SHE 作为选题策略；(3)使用极大似然(maximum likelihood estimation, MLE)方法估计被试的属性掌握模式。"。详见文章第 15 页。

意见 10：第 16 页：（1）第 1 行，"估计正确率"读起来别扭，建议使用"估计精度"；（2）RMSE 中的"估计正确性"也建议改为"估计精度"；（3）"4.4 研究 1 结果"第 3 行，"…以及均值分别为…"表述不准确；（4）"SIE 方法的 AVCER 值为 0.0%"，SIE 方法既然一次都判不准，那么为何将它作为比较基准？文中如果要呈现 SIE 的结果，是否可以考虑改进 SIE 方法在 G-DINA 模型下的不足？

回应：非常感谢您的建议！（1）我们已将 AVCER 中的"估计正确率"和 RMSE 中的"估计正确性"均修改为"估计精度"。

（2）我们将"…以及均值分别为…"这整个句子修改为"各模拟条件下 SCADOCM 的平均运行时间(ART)、属性向量正确估计率(AVCER)以及均方根误差(RMSE)的均值分别为 5.231s、66.4%和 0.102，SIE 方法对应的值分别为 99.893s、0.0%和 0.242。"。

（3）SIE 方法在 DINA 模型下具有较好的性能(Chen et al. 2015)，但 G-DINA 模型下的 AVCER 值几乎都为 0.0%，该结果可以为其他研究者和实践者提供参考与借鉴，他们未来在 G-DINA 等饱和模型下进行在线标定方法研究时可以避免选择该方法作为比较基准。因此，研究 1 在 SIE 方法的 AVCER 极低的情况下保留了该方法作为比较基准，但研究 2 中未使用该方法。此外，我们也在文章中给出了 SIE 方法的 AVCER 值极低的原因，并对 SIE 方法提出了未来改进的措施。具体来说，使用 SIE 标定新题 *Q* 矩阵时，基于模型复杂性的考虑，对似然进行惩罚，构建 BIC 指标，选择能使 BIC 值最小的 q 向量作为新题的估计 q 向量。我们初步的预实验表明：改进的 SIE 方法的项目标定精度优于 SIE 方法。项目参数 $P(\mathbf{0})$和 $1\text{-}P(\mathbf{1})$的取值范围为 $U(0.1, 0.3)$，属性掌握模式分布为正态分布，标定样本为 500 时，改进 SIE 方法的平均运行时间(ART)、属性向量正确估计率(AVCER)、项目参数均方根误差(RMSE)、$P(\mathbf{0})$和 $1\text{-}P(\mathbf{1})$参数的 RMSE 值分别为 153.758s、54.9%、0.104、0.058 和 0.048，*Q* 矩阵标定精度远优于 SIE 方法，但仍不如本文提出的新方法(此条件下 SCADOCM 的 AVCER 值为 61.7%)。我们已在"讨论与未来研究方向"中增加该内容。

意见 11：图 1~3 中，项目质量 U(0.1-0.3)和 U(0.05-0.15)的表述方式不准确。

回应：非常感谢您的宝贵建议！我们已对图 1~3 进行修改。

意见 12：第 19 页第 3 行，“SIE 方法…AVCER 值均接近于 0”与前面的描述不一致，前面的描述是“SIE 方法的 AVCER 值为 0.0%”。
回应：非常感谢您为我们指出的问题！SIE 方法的 AVCER 值均接近于 0.0%，我们已将前后的表述修改为一致。

意见 13：除了呈现结果，必要时还需要提供解释。比如：（1）第 19 页“SCADOCM 的 Q 矩阵估计精度在属性模式为均匀分布时最好，高阶分布时次之，正态分布时最差”，为什么会出现这个结果？（2）第 21 页“SCADOCM 和 SIE 的项目参数标定精度…，在部分实验条件下随项目质量的提升而略有下降”，为什么项目质量更好了标定精度反而会略有下降？
回应：非常感谢您的宝贵建议！我们在呈现结果时增加了相应的解释。（1）SCADOCM 的 *Q* 矩阵估计精度在属性模式为均匀分布时最好，高阶分布时次之，正态分布时最差。其可能的原因在于，均匀分布下每种属性掌握模式的被试人数都较为均匀，而高阶分布和正态分布下某些属性掌握模式的被试人数非常少，尤其是正态分布下某些属性掌握模式的被试人数更少，这不利于正确 *q* 向量的识别(Chiu, 2013; Wang et al., 2018)，从而导致高阶分布和正态分布下的 *Q* 矩阵估计精度更低。

（2）“SCADOCM 和 SIE 的项目参数标定精度…，在部分实验条件下随项目质量的提升而略有下降”。具体来说，在属性掌握模式为正态分布，尤其是标定样本少的情况下，SCADOCM 和 SIE 的项目参数标定精度随项目质量的提升而略有下降。在标定样本为 100 时，SCADOCM 在两项目参数范围间的 RMSE 差值最大，达到 0.014；在标定样本为 50 时，SIE 在两项目参数范围间的 RMSE 差值最大，达到 0.019。这可能是标定样本和属性掌握模式分布相互作用的结果。新题的项目参数标定精度在标定样本量少的情况下较低，而在标定样本少且属性掌握模式分布为正态分布时，更是有可能出现某些属性掌握模式下的被试数量多而另一些属性掌握模式下的被试缺失的情况，两者共同作用可能导致项目质量高时的 RMSE 值略大于项目质量低时，但是这种差异是较小的，且可以通过增大样本量或改变属性掌握模式分布扭转这种趋势。

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598-618.

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement*, *42*, 446–459.

意见 14：第 24 页表 2 中只呈现了两个参数的描述性统计结果，为什么不呈现其他参数的结果？
回应：非常感谢您的宝贵建议！网络成瘾题库测量 9 个症状标准，共 512 种属性掌握模式，如果在表格中列出所有属性掌握模式的结果，表格过大，显得累赘。因此，文章中仅给出了能较好地体现整个题库项目参数的范围的 1-*P*(**1**)和 *P*(**0**)的结果。

意见 15：第 26 页：“SIE 在标定样本少、属性掌握模式分布为均匀分布和高阶分布时的标定效率更高…”。这里涉及“标定样本少”等多个条件，那么 SIE 是相对什么条件的标定效率会更高？下同。另外，文中还有一些错别字，比如“的地得”不分（如第 10 页“较好的体现了”）。
回应：非常感谢专家的建议！我们将“SIE 在标定样本少、属性掌握模式分布为均匀分布和高阶分布时的标定效率更高…”修改为“SIE 方法在标定样本少时的标定效率比标定样本大时更高，在属性掌握模式分布为均匀分布和高阶分布时的标定效率比属性掌握模式分布为正

态分布时更高”，使表述更为清楚。此外，我们认真仔细地阅读了文章，修改错别字，提高文章质量。

……

审稿人 2 意见：

本文提出一种基于 GDINA 模型的 CDCAT 在线标定方法。从研究进展上看，无论国内还是国外，在 CDCAT 中基于 GDINA 进行新题目标定的研究还很少，因此本研究具有一定的前沿性；从理论上讲，本研究具有一定的理论创新，尝试将 SCAD 引入在线标定中；从实践角度看，本研究属于“空中楼阁”式研究，作者未展现出的具体的落地应用场景。整体而言，本研究属于缺乏明确心理学问题的算法或技术工作。目前中国心理学的大多数研究者和《心理学报》读者群体的数理水平仍不高，难以理解本研究的创新性、贡献和价值。尽管个人建议作者主动将本文修改为英文，投稿国际更专业的期刊，如 APM、JEM 等，以便研究成果获得更多关注，但仍由主编等做出最终决定。

意见 1：引言部分，使用 DCM 建立题库，对新题进行标定的必要性是什么？与连续特质测量模型不同，DCM 中类别特质的“量尺”不是随意的(Madsion & Bradshaw, 2018, https://doi.org/10.1007/s11336-018-9638-5)，并不需要通过固定题目参数的“量尺”来保证被试参数估计值在一个量尺上。因此，从文章的完备性上，建议作者进一步补充 DCM 中对题目进行标定的必要性以及作用。

回应：非常感谢您的宝贵意见！我们已在文章引言部分增加了 DCM 中对题目进行标定的必要性以及作用，具体如下：

“CD-CAT 中可使用在线标定技术标定新题的参数，但有一个问题值得思考，即认知诊断测验中是否需要进行等值，是否有必要使用在线标定技术对新题进行标定？de la Torre 和 Lee (2010)在研究中指出当模型与数据完全拟合时，决定型输入噪音与门(the deterministic input, noisy and gate, DINA; Junker & Sijtsma, 2001)模型的项目参数具有不变性；Bradshaw 和 Madsion (2015)，Madsion 和 Bradshaw (2018)也在其研究中指出对数线性认知诊断模型（log-linear cognitive diagnosis model CDM, LCDM; Henson et al., 2009）和基于 LCDM 模型开发的 TDCM (the Transition Diagnostic Classification Model)在模型与数据拟合的情况下参数具有不变性。在此条件下，无需通过等值来保证被试参数估计值在同一量尺上。然而，其研究也指出在模型与数据不完全拟合时，参数不变性不成立；且即使模型与数据拟合的情况下，参数不变性也会随着标定样本的减少而减弱(Bradshaw & Madsion, 2015; de la Torre & Lee, 2010; Madsion & Bradshaw, 2018)。这表明参数不变性成立需满足一些必备的条件：如模型与数据完全拟合，标定样本量足够大(如不少于 1000)，在这些条件下可以不用进行等值。但在实际测验情境中，模型与数据完全拟合的情况并不总能得到满足，且在同一次测验中也较难获得足够大的标定样本，这都会导致项目参数估计出现偏差，进而影响被试的分类准确性和 *Q* 矩阵的标定正确性。因此，在 CD-CAT 题库建设中有必要进行在线标定，这有利于降低项目参数估计偏差所带来的影响，提高 CD-CAT 题库和测验的质量。”。

Bradshaw, L. P., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, *16*(2), 99-118.

de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*, 115–127.

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

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*, 963-990.

意见 2：引言部分，DINA 模型是否适用于 DSM-V，跟成瘾诊断标准有什么关系？GDINA 如何适用于 DSM-V 的诊断标准？

回应：非常感谢您的宝贵意见!DSM-V 中界定的成瘾诊断标准，共有 9 条症状标准(即属性)，被试符合其中 5 条及 5 条以上症状可诊断为网络成瘾。对于 DINA 模型，它假定被试在项目上的作答只受到项目测量的所有属性的交互作用影响，而不受主效应及其它类型的交互作用的影响；而 G-DINA 模型则没有这些严格的假设，认为被试的作答可以是由项目测量的各属性的主效应与各种类型的交互效应的共同影响，如果主效应（或交互效应）的系数估计值为 0 或接近 0，则这时主效应（或交互效应）的作用不明显，即这时不存在主效应（或交互效应），但若系数显著不为 0，则说明存在主效应（或交互效应），因此 G-DINA 模型更为灵活。当我们在 G-DINA 模型的基础上假设只存在所有属性的交互作用，那这时 G-DINA 模型就简化为 DINA 模型，因此 DINA 模型是 G-DINA 模型一个特例。在 DSM-V 关于网络成瘾诊断标准中，它有 9 条症状标准，也即 CDM 中的 9 个属性，我们可以通过 CDM 来估计被试在 9 条症状标准上的情况（可以是 0-1 的表示，即是否具备该症状；也可以是连续的表示，即具备该症状的概率），并结合 DMS-5 诊断标准（如被试符合其中 5 条及 5 条以上症状可诊断为网络成瘾）实现对被试的诊断。对于本研究，我们采用的是假设条件更少的饱和认知诊断模型(G-DINA 模型)进行数据分析，我们也补充了相关模型-资料拟合检验的结果(文章中表 4)，发现 G-DINA 模型较 DINA 等其它约束的认知诊断模型更能拟合该网络成瘾数据。

表 4 网络成瘾题库模型-资料拟合检验结果

模型	AIC	BIC	LL
DINA	309348.5428	314897.6939	-153637.2714
DINO	309803.4409	315352.5920	-153864.7204
ACDM	307764.2211	313586.2812	-152794.1105
G-DINA	307426.2025	313574.6833	-152564.1012

意见 3：引言部分，本研究的具体心理学问题是什么？针对心理学、教育学或行为科学领域读者，CDCAT 的应用价值是什么？CDCAT 中研究题目标定的必要性是什么？是否有实践动机？

回应：非常感谢您的建议！（1）本研究的具体心理学问题是针对 CD-CAT 题库开发与维护过程中项目增补的技术难点，开发高效可行的在线标定方法。尽管研究是开发新的方法，但我们认为其与心理学问题是紧密相关的。心理测量学是我们研究心理学的工具，心理问题(如抑郁、焦虑)的评估与测量都离不开心理测量学。CD-CAT 作为一种新的测验形式，可以更高效、精准地筛查存在心理问题的患者，缓解患者(如抑郁症、躁狂症)做包含大量题目的问卷时的痛苦，减轻其测试的负担。更为重要的是，CD-CAT 可以帮助测验使用者了解患者在某种心理问题各个症状上的表现，更快地获得诊断结果，且能依据该诊断结果制定针对性的治疗方案。在心理测评中应用 CD-CAT 对患者和测验使用者都具有重要的意义，但 CD-CAT 在心理与教育测评实践中的应用受到了诸多限制，题库的构建和维护是其在实际测验中难以被应用的主要原因之一。本研究致力于解决题库构建与维护过程中进行项目增补所面临的技术重点和难点，促进 CD-CAT 在心理测评实践中的应用与推广，以期帮助测验使用者获得更为精细的诊断结果，制定相应的治疗计划，这与心理学问题息息相关。

（2）CD-CAT 的应用价值：CD-CAT 通过“量体裁衣”的个性化测验快速准确地探查被

试在所测内容上的优势和不足，可及时为被试提供精细的诊断反馈信息，在提高测验结果准确性的同时极大地减轻了测验参与者的作答负担(Lin & Chang, 2019)。在心理测量中，如果测验能快速、准确、高效地为临床心理医生提供来访者在某一心理问题上的具体症状表现，那么心理医生可结合具体症状及时采取相应的治疗方法，有助于提升心理治疗的效果。而在教育测评中，如果测验能快速、准确、高效地为教师提供学生掌握和欠缺的具体知识点，那么教师在课堂上可以重点讲授学生有待提高的知识点，学生也可以针对自己的弱项进行有针对性的学习，从而减轻学生负担，改进教学，提高教学效果。这符合“双减”等政策的精神和要求，也较好地满足当了前国家和社会发展的实际需要，有利于促进精准、自适应和个性化的心理与教育测评，以及考试的数字化革新。

有关 CD-CAT 中研究题目标定的必要性问题，我们已在问题 1 中进行了回答。

研究的实践动机在于：基于研究 2 的网络成瘾题库构建网络成瘾 CD-CAT 在线测试平台，将 SCADOCM 等在线标定方法用于该测验题库的维护，以保证该测验能持续用于网络成瘾患者高效、精准地筛查，了解其具体的症状特征，为后续的治疗提供建议与参考。总而言之，本研究的实践动机在于解决 CD-CAT 应用于实践测验时所面临的重要挑战之一，即 CD-CAT 题库的维护，促进 CD-CAT 在实践中的应用与推广，更好地为心理与教育测评服务。

根据专家的意见，我们在文章引言和讨论部分对上述内容了作了进一步的补充，感谢审稿人的这条宝贵意见。

Lin, C. J., & Chang, H. H. (2019). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educational and psychological measurement*, *79*(2), 331657.

意见 4：方法部分，部分公式中参数的含义为给予说明，比如，公式 3 中的 U_io。

回应：非常感谢您的建议，我们已仔细核对文章公式，新增了部分公式中参数的含义说明。

意见 5：研究 1 部分，数据生成。如何生成新题目的作答数据？是否有研究关注 CAT 中不同选择策略对新题的标定结果的影响？作者为何使用 SHE 选题策略？

回应：非常专家的建议！（1）新题作答数据的生成：与旧题作答数据的生成一致，基于模拟的项目参数真值及被试属性掌握模式真值，根据认知诊断模型计算被试在每个新题上的正确作答概率，将该正确作答概率与 0~1 之间的随机数进行比较，如果被试在题目上的正确作答概率大于随机数，则答对题目，否则答错题目。

（2）已有研究者探索过选题策略对新题标定效果的影响。Chen 等人(2015)在研究中比较了随机选题策略和 SHE 选题策略对新题标定结果的影响，结果表明 SHE 选题策略下的新题标定精度优于随机选题；谭青蓉等人(2021)在其研究中比较了 PWKL、MPWKL、SHE 和 GDI 选题策略下新题的标定精度，整体上 SHE、MPWKL 和 GDI 选题策略具有相似的 *Q* 矩阵标定精度，略优于 PWKL 选题策略；Tan 等人(2022)的研究也发现真实题库下 SHE 选题策略下的 *Q* 矩阵标定精度略优于 PWKL 选题策略。综上，在选题策略为 SHE 时可以获得较好的新题标定精度，且 SHE 选题策略理论基础扎实，具有较高的估计精度(张学工，2010; Zheng & Chang, 2016)。此外，已有同时标定新题 *Q* 矩阵和项目参数的研究中，大多使用 SHE 选题策略。因此，研究中选用 SHE 作为选题策略。

我们已在文章中增加了新题作答数据的生成，以及研究中选用 SHE 作为选题策略的理由。

Chen, Y., Liu, J., & Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. *Applied psychological measurement*, *39*(1), 5-15.

Tan, Q., Cai, Y., Luo, F., & Tu, D. (2022). Development of a High-Accuracy and Effective Online Calibration Method in CD-CAT Based on Gini Index. *Journal of Educational and Behavioral Statistics*, *48*(1), 103-141.

Zheng, C., & Chang, H. H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*(8), 608-624.

谭青蓉, 汪大勋, 罗芬, 蔡艳, 涂冬波. (2021). 一种高效的 CD-CAT 在线标定新方法：基于熵的信息增益与 EM 视角. *心理学报*, *53*(11), 1286-1300.

张学工. (2010). *模式识别(第三版)*. 清华大学出版社.

意见 6：研究 1 部分，RMSE 指标过于笼统，是 GDINA 中所有题目参数估计精度的笼统反映，建议至少拆解为 guess 和 slip 两个上下限参数。

回应：非常感谢您的建议！我们已在研究一中增加了 guess ($P(\mathbf{0})$)和 slip (1- $P(\mathbf{1})$)的结果。详见文章第 21、22 和 23 页。

意见 7：研究 1 部分，统一几个图中两种方法所使用的颜色。

回应：非常感谢您的宝贵意见！正文三个图中，我们统一使用绿色表示 SIE 方法，红色表示 SCADOCM 方法。

意见 8：研究 1 部分，作者需解释图 2 中 SIE 的结果。逻辑上讲即便随机分配也不会是 0 左右的判准率。

回应：非常感谢您的宝贵意见！图 2 结果表明，SIE 的属性向量估计正确率(AVCER)在各条件下几乎都接近于 0。其可能的原因在于：研究中所用评估 Q 矩阵估计精度的 AVCER 指标，评估题目的整个估计 q 向量和真实 q 向量之间的一致性，也即 q 向量模式的估计精度。SIE 方法中使用 MLE 方法估计新题 q 向量，而在 G-DINA 模型下，MLE 方法倾向于选择测量所有属性的 q 向量（即全为 1 的 q 向量）作为新题的估计 q 向量(汪大勋 等, 2020; Chen et al., 2013)。例如，测验测量属性个数 $K = 5$ 时，SIE 方法选择 q 向量 q=[1 1 1 1 1]作为题目的估计 q 向量，实验结果调查也证实了这一点。在模拟实验中，设置测验共测量 5 个属性，每个题目(旧题和新题)最多测量 3 个属性，使用 SIE 标定新题 Q 矩阵偏向于指定每个题目都测量 5 个属性，此时新题 Q 矩阵的属性向量估计精度低于随机分配概率，出现 AVCER 在 0 左右的结果。假设 20 个新题均测量 3 个属性，则20×5的新题 Q 矩阵中有 60 个元素为 1，40 个元素为 0，此时 SIE 方法的属性估计精度约为 60%，也即 SIE 方法的属性估计精度最大值为 60%；研究中 20 个新题的 q 向量从 300 个旧题(测量 1、2 和 3 个属性的项目均为 100 题)中随机抽取，20×5的新题 Q 矩阵中元素为 1 的个数大多数情况下小于 50 个，该类情况下 SIE 方法的属性估计精度低于 50%。研究 1 中各模拟条件下 SIE 方法的平均属性估计精度为 39.8%，大于 0，低于 50%。我们已在“讨论与未来研究方向”部分增加了该解释。

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140.

Wang, D., Gao, X., Cai, Y., & Dongbo, T. U. (2020). A method of q-matrix validation for polytomous response cognitive diagnosis model based on relative fit statistics. *Acta Psychologica Sinica*, *52*(1), 93-106.

意见 9：研究 2 部分，存在的必要性不高，没有给出额外的结果。通过在研究 1 中调控题目参数分布即可实现。

回应：非常感谢您的建议。研究 1 是基于模拟题库的研究，可以通过调控题目参数分布来获得与研究 2 较为相似的分布的题库，但我们认为两者之间还是存在不同的。具体来说，研究 2 中真实题库的项目参数是基于真实测验中被试的作答数据估计获得，是真实情境中各种因素综合影响下的结果，而研究 1 中题库 Q 矩阵和项目参数的模拟是在一种理想的状态下进

行，不论怎么调控题目参数分布，都很难综合考虑到真实测验情境中可能存在的各种影响因素，与真实的题目参数之间存在差异。真实题库的项目参数分布可能较难通过直接模拟获得，即使模拟的项目参数分布与真实题库的项目参数分布相似，两者也很难完全一致。

因此，我们进行了研究 2，更多地是想探讨在真实题库而非我们模拟的题库下新方法 SCADOCM 的性能，保证研究的生态性。同时，研究 2 结果进一步说明 SCADOCM 方法的标定效果即使在真实题库下仍较理想，可推广性较好，可以为实践应用者提供更贴切、具体的参考，有保留的价值。

意见 10：研究 2 部分，如果研究 2 是基于实证研究中的题目库，参数的“真值”是什么？是已有研究中 GDINA 的估计结果？似乎只能反映一致性，而不是准确性。

回应：非常感谢您的宝贵建议。研究二中参数的“真值”是基于已有研究中给定的由专家标定的 Q 矩阵和所有被试的真实作答数据使用 G-DINA 模型估计的结果。如您所提到的，RMSE 反映了项目参数估计结果间的一致性，而非准确性，我们已将研究 2 中有关项目参数“准确性”的表达修改为“一致性”。

意见 11：其他，缺乏实证研究情境。

回应：非常感谢您的宝贵建议。与以往国内外项目参数同时性在线标定方法的研究（Chen et al., 2015; Tan et al., 2022; 陈平, 辛涛, 2011; 谭青蓉 等, 2021）一致，本研究仅在真实和模拟的题库下检验了新方法的性能，并未在实证研究情境中加以应用，评估其性能。主要原因在于：在真实测验情境中验证在线标定方法的性能，需要事先构建好一个可以用于实际测验的真实 CD-CAT 测试平台，这需要耗费大量的时间和精力，目前这种平台我们较难获取。这是本研究，甚至于目前 CD-CAT 中在线标定研究的不足之处，也是未来可进一步深入的研究方向。我们已在“讨论与未来研究方向”中增加了该讨论。

最后，再次感谢审稿专家的细致审稿，并为本文进一步完善提出了非常宝贵的修改意见！

第二轮

审稿人 1 意见：

经过作者修改，文章质量得到很大提升，已达到发表水平。目前版本的稿件中还有一个小地方需要修改，如下：

图 1~3 中的“U(0.1-0.3)和 U(0.05-0.15)”改为“U(0.1, 0.3)和 U(0.05, 0.15)”。

回应：非常感谢您的宝贵意见！我们已将图 1~3 中的“U(0.1-0.3)和 U(0.05-0.15)”改为“U(0.1, 0.3)和 U(0.05, 0.15)”。

审稿人 2 意见：

意见 1：题目与正文匹配程度有待提高，题目并没有体现出 GDINA；

回应：非常感谢专家的宝贵意见！根据您的建议，我们将题目修改为：“CD-CAT 中基于 SCAD 惩罚和 EM 视角的在线标定方法开发——基于 G-DINA 模型”。

意见 2：引言第一段内容缺少具体参考文献，比如，Tang 和 Zhan(2021; doi: 10.1177/23328584211060804)表明诊断反馈可以有效促进学习，那是否有研究表明“心理医生

可结合具体症状及时采取相应的治疗方法，有助于提升心理治疗的效果”？

回应：非常感谢专家的宝贵意见！我们已补充了相关文献。另，有研究者指出 CDM 可以提供心理评估与干预中更精细的证据，帮助临床医生更好地理解心理问题及一些具体症状之间潜在的复杂关系，心理医生可及时地制定有效的预防和干预策略，推进心理治疗进程(如, de la Torre, 2018; Tan et al., 2023)。我们在引言第一段增加了相应文献，并进行了修改，使表达更清晰。具体如下：

“在心理评估中，如果测验能快速、准确、高效地为临床心理医生尤其是新手医生提供来访者在某一心理问题上的具体症状表现，帮助临床医生更好地理解心理问题及一些具体症状之间潜在的复杂关系，心理医生可及时地制定有效的预防和干预策略，推进心理治疗进程(如, de la Torre et al., 2018; Tan et al., 2023)。而在教育测评中，如果测验能快速、准确、高效地为教师提供学生掌握和欠缺的具体知识点，教师在课堂上可以重点讲授学生有待提高的知识点，学生也可以针对自己的弱项进行有针对性的学习，从而减轻学生负担，改进教学，提高教学效果(如，Tang & Zhan, 2021)。”。

de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, *51*(4), 281-296.

Tan, Z., de La Torre, J., Ma, W., Huh, D., Larimer, M. E., & Mun, E.-Y. (2023). A tutorial on cognitive diagnosis modeling for characterizing mental health symptom profiles using existing item responses. *Prevention Science: The Official Journal of the Society for Prevention Research*, *24*(3), 480-492.

Tang, F., & Zhan, P. (2021). Does diagnostic feedback promote learning? Evidence from a longitudinal cognitive diagnostic assessment. *AERA Open*, *7*(3), 296-307.

详见文章第 1 页。

意见 3：引言中“这表明参数不变性成立需满足一些必备的条件：如模型与数据完全拟合，标定样本量足够大(如不少于 1000)，在这些条件下可以不用进行等值。”该已有研究结论与本文研究设计之间的关系是什么？本文研究中数据生成数据与分析数据完全一样，是否有必要进行等值或标定？作者设定的被试量有大于 1000 有小于 1000，结果的差异性是否体现了小于 1000 人需要进行等值（标定）？

回应：非常感谢您的宝贵意见！我们在引言中提到上述已有研究结论，主要是为了说明测量不变性需要在一定的条件下成立，而这些条件在实际测验中并不总是能得到满足，因而在 CD-CAT 题库的构建与维护过程进行等值或标定是有必要的。因此，本文参考已有 CD-CAT 在线标定研究(如，Chen et al., 2012; Chen et al., 2015; Tan et al., 2022; 陈平, 辛涛, 2011; 谭青蓉 et al., 2021)的研究设计，围绕在线标定方法的性能验证及相关因素对其的影响进行研究设计，而未结合该结论进行研究设计。

此外，为进一步说明测量不变性的问题，我们拟通过一个模拟实验检验 CD-CAT 情境中，不进行在线标定时新题的 Q 矩阵估计精度和项目参数估计精度。无标定(不进行在线标定)主要指仅基于 CD-CAT 过程中所收集到的被试在新题上的作答数据估计新题 Q 矩阵与项目参数，而不利用在 CD-CAT 过程中所获得的被试属性掌握模式等相关信息（这些信息是锚定新题参数量尺或实现参数等值的关键信息）。模拟实验中，属性掌握模式分布为均匀分布，项目参数范围为 $U(0.1, 0.3)$，标定样本为 500、1000 和 2000。除是否进行标定这一不同之处外，模拟过程与研究 1 保持一致。实验结果如下表 1 所示：相比于标定条件，SIE 方法和 SCADOCM 方法的 Q 矩阵估计精度和项目参数估计精度在无标定条件下均较低。这表明即使模型数据完全拟合(即产生数据的模型与估计数据的模型一致)，标定样本足够大(如 1000、2000)的情况下，在 CD-CAT 中往题库中增补新的项目时，进行等值/标定是有必要的。这与以往基于模拟数据的研究结论不一致，其原因可能在于：以往研究在题目 Q 矩阵已知

且正确的情况下，基于完整的被试作答矩阵估计新题的项目参数。在此条件下，在样本量大(如 1000)时可以获得较高的被试分类精度和项目参数估计精度，观察到测量不变性；但不同于以往研究，本研究模拟 CD-CAT 情境下新题 Q 矩阵与项目参数的标定，此时被试的作答矩阵是一个缺乏较多作答数据的稀疏矩阵，每个题目都只有部分被试作答，每个被试也只作答少数几个题目(若被试需作答的待标定新题过多，CD-CAT 的测验长度可能大幅增加，加重被试的作答负担)，且题目 Q 矩阵未知。此时，即使标定样本大，项目参数的标定精度也较低(如下表 1 所示)，可能难以保证测量不变性。Bradshaw 和 Madsion (2015)在其研究中指出，在参数估计精度较低的情况下，很难观察到较强的测量不变性，其在研究中也提到，模型数据拟合假设以其它形式违背(如 Q 矩阵错误指定，Bradshaw & Madsion, 2015)时，可能也会影响被试的分类一致性。后续研究可以探索 Q 矩阵指定错误和稀疏作答矩阵等因素对测量不变性的影响。为此，我们在文章的讨论部分，补充了上述相关内容，再次感谢审稿专家的宝贵意见。

表 1 无标定和标定情况下各在线标定方法的项目标定结果

标定样本	是否标定	SIE				SCADOCM			
		AVCER(%)	RMSE	P(0)	P(1)	AVCER(%)	RMSE	P(0)	P(1)
500	无标定	0	0.474	0.267	0.216	6	0.402	0.155	0.157
	标定	0	0.119	0.111	0.113	72.7	0.055	0.068	0.046
1000	无标定	0	0.474	0.207	0.216	7	0.405	0.159	0.173
	标定	0	0.083	0.079	0.078	82.55	0.037	0.041	0.032
2000	无标定	0	0.468	0.207	0.216	5	0.441	0.275	0.276
	标定	0	0.058	0.053	0.054	89.35	0.025	0.027	0.022

鉴于文章篇幅过长，我们仅在回复中呈现了如上补充实验及其结果，供审稿专家参考。

意见 4：核对公式与结果所用公式之间的匹配性，比如 RMSE；

回应：非常感谢您的宝贵意见！我们认真核对了文章中的公式和结果所用公式之间的匹配性，结果均是基于文章中所呈现的公式计算。此外，我们增加了 $P(\mathbf{0})$和 $P(\mathbf{1})$参数的 RMSE 计算公式，具体如下：

“此外，$P(\mathbf{0})$和 1-$P(\mathbf{1})$参数的 RMSE 计算公式与公式(19)略有不同，具体如下所示：

$P(\mathbf{0})$：
$$RMSE=\sqrt{\frac{1}{100\times m}\sum_{r=1}^{100}\sum_{j=1}^{m}\left(\hat{P}_j^{(r)}(0)-P_j^{(r)}(0)\right)^2}, \tag{20}$$

1-$P(\mathbf{1})$：
$$RMSE=\sqrt{\frac{1}{100\times m}\sum_{r=1}^{100}\sum_{j=1}^{m}\left(\left(1-\hat{P}_j^{(r)}(1)\right)-\left(1-P_j^{(r)}(1)\right)\right)^2}. \tag{21}$$”

详见文章第 16 页。

意见 5：如果作者执意要保留研究 2，请将表 3 替换为全信息，即所有题目的参数值或分布，而非一个笼统的描述性统计。

回应：非常感谢您的建议！我们保留了表 3，并根据您的建议，在文章中增加了题库中所有题目的参数值，但考虑到数据表过大，我们将该结果放在附录(附表 1)中。

详见文章第 40 页。

最后，再次感谢审稿专家的细心审阅，感谢您们为本文进一步完善提出了非常宝贵的修改意见！

第三轮

审稿人 2 意见：

经过修改，文章质量得到进一步提升。但仍存在一些语义不清的地方，希望作者能更加严谨地阐述学术观点。比如，

意见 1：第 6 页，“如 Liu 等人(2013)开发的中国大型英语二级测验题库，项目参数范围在 0.001 到 0.5 之间。”具体是指什么题目参数？

回应：非常感谢专家的宝贵意见！此处的项目参数指项目失误参数，即被试掌握了项目测量的所有属性但错误作答该项目的概率。为使表述更为清晰、严谨，我们已对此进行修改，具体如下：

“如 Liu 等人(2013)开发的中国大型英语二级测验题库，其项目失误参数(被试掌握了项目测量的所有属性但错误作答该项目的概率)的范围在 0.001 到 0.5 之间。”。

意见 2：第 6 页，“但其方法受被试分组数量的影响，其在 G-DINA 等模型下的性能可能并不理想。”其中，“被试分组”具体是指什么？

回应：非常感谢专家提出的宝贵修改意见。“被试分组”具体指被试的类别，如 DINA 模型在每个项目上将被试区分为两个类别，G-DINA 模型在每个项目上将被试区分为$2^{K_j^*}$(K_j^*表示项目 j 测量的属性个数)个类别。我们对此进行了阐明，以便能更准确地表达，具体如下：

“但该类方法受被试类别数量的影响，DINA 模型在每个项目上均将被试区分为两个类别，而 G-DINA 模型在每个项目上将被试区分为$2^{K_j^*}$(K_j^*表示项目 j 测量的属性个数)个类别，其在 G-DINA 等模型下的性能可能并不理想。”。

意见 3：不限于上述两点，请作者通读全文。

回应：非常感谢专家提出的宝贵修改意见。我们对全文进行了仔细的核查，对存在语义不清的内容进行了修改，以使文章表述更为清晰、严谨。

最后，再次感谢审稿专家细致、严谨的审稿工作，为本文进一步完善提出了非常宝贵的修改意见！

第四轮

审稿人 3 意见：

我认真阅读了三轮的修改意见、作者回复以及论文文稿，意见予以接收发表。研究提出一种适用于 G-DINA 等模型的同时标定新题 Q 矩阵与项目参数的认知诊断计算机化自适应测验(CD-CAT)在线标定新方法 SCADOCM，在研究设计、模拟结果、以及实证数据等方面都展示了本方法的可行性与优越性。研究选题略显狭窄，只是拓展到了 GDINA 模型，稍微降低了本文的影响力。不过，基于以上全面的考虑，本人建议发表。

编委意见：这篇论文经过了很长时间的修改，鉴于论文提出一种适用于 G-DINA 等模型的

同时标定新题 Q 矩阵与项目参数的认知诊断计算机化自适应测验(CD-CAT)在线标定新方法SCADOCM，在研究设计、模拟结果、以及实证数据等方面都展示了本方法的可行性与优越性，加上作者不断的修改完善。我建议发表。

主编意见：建议将结论从讨论中提炼出来，单独写成一部分，以方便读者阅读。
回应：非常感谢主编专家对稿件提出的宝贵意见。为方便读者阅读，我们已根据主编的建议，将结论从讨论中提炼出来，单独写为一部分。