《心理学报》审稿意见与作者回应

题目：基于选项层面的认知诊断非参数方法
作者：郭磊，周文杰

第一轮

审稿人 1 意见：

认知诊断评估作为当前心理测量领域较为先进的测评框架，可以对学生在各知识内容上的掌握情况进行较为精细的诊断，这为诊断教学和个性化学习提供了相应可能，因而受到研究者和教育工作者的青睐。论文《基于选项层面的认知诊断非参数方法》（xb20-447）使用非参数诊断分类方法将选择题中的干扰项信息给予考虑，提出了三种适用于能够处理干扰项诊断信息的非参数认知诊断方法，该方法不受样本量的影响，这为小样本的课堂班级评估提供了条件，因而具有较大的实践意义。作者先详细地介绍了该非参数方法的操作流程，随后通过两个模拟研究和一个实证研究对所提方法的性能进行了较为系统的验证。论文逻辑清晰，结构合理，内容也相对厚实，可读性相对较强，具有发表的价值。但论文还存在一些需要进一步修改的地方，有以下几点：

意见 1：第 2 章节，据审稿人所知，开放题通常不存在猜测行为，如此针对开放性题目设置相应的权重是否有必要？作者应进一步明确惩罚汉明矩阵方法的应用情境。

回应：非常感谢审稿人的提问。Chiu 和 Douglas（2013）首次提出了在传统 0-1 计分测验中的惩罚汉明方法，其中便介绍了惩罚汉方法的应用情景。以开放题为例，被试作答开放题时几乎不可能发生猜测行为，换言之，猜测行为发生的概率很小，所以当猜测行为发生时，即理想作答向量（ideal response vector）为 0 而实际作答向量为 1 时，应给予该情况更高的权重，以使该理想反应模式与实际作答的距离增大，符合现实情景。正因为开放题的猜测行为发生的概率很小，而失误行为发生的概率相对较大，所以需要为猜测行为设置相比于失误行为更大的权重以达到“惩罚”猜测行为的目的。修改稿中我们已在第二稿第 2 章节补充说明其使用场景。

而在 MC 测验中，猜测行为和失误行为的发生概率不存在明显的差异，但不同题目间可能存在较大的质量差异，会导致不同题目间的失误参数存在较大差异，反映到被试作答行为上，则是被试在不同质量题目间发生失误行为的概率可能会存在较大差异。本文针对 MC 测验的这一特点提出的适用于 MC 测验的惩罚汉明距离方法（d_{ph-MC}距离法），本质上是为不同质量的题目设置不同的权重，以“惩罚”在失误参数较小的题目发生失误行为的 KS，即本不应该发生失误，但却发生了，所以要给予较大“惩罚”。例如，假设选择题 1 和选择题 2 有相同的题目 q 向量，A 选项为正确答案，如下表：

选择题 1 和选择题 2 的 q 向量示例

选项	属性 1	属性 2	属性 3	属性 4	属性 5
A	1	1	0	1	0
B	1	1	0	0	0
C	1	0	0	0	0
D	0	0	0	0	0

选择题 1 的失误参数为 0.1，选择题 2 的失误参数为 0.4。当 KS 为(1, 1, 0, 1, 0)，此时的 KS

与 A 选项所考察的属性均相匹配，选择题 1 的失误参数更小，意味着该 KS 更不可能在选择题 1 上发生失误行为去选择其他选项，选择题 1 的鉴别正常作答和失误行为的能力更强。因此，应给予选择题 1 更高的失误权重，以惩罚在该题目发生失误行为的 KS。
综上所述，本文提出的d_{ph-MC}距离法的逻辑及应用场景与传统的惩罚汉明方法有所不同，是对传统方法的改进。传统惩罚汉明方法应用于某些题型的失误行为与猜测行为发生的可能性存在较大差异，而d_{ph-MC}距离法的主要应用场景是 MC 测验中不同题目间的质量存在较大差异。

意见 2：第 3.3 章节，作者介绍了惩罚汉明距离。审稿人的疑惑是‘非参数方法难以知晓题目质量，此时如何设置相应的惩罚权重？’
回应：非常感谢审稿人的提问。Chiu 和 Douglas（2013）提出的传统 0-1 计分测验中的惩罚汉明距离方法并未给出惩罚权重的具体的设置方法，只是给定：当$s > g$时，$w_s < w_g$；当$g > s$时，$w_g < w_s$。我们推测，Chiu 的论文中的权重取值应该是通过探索得到的。在本文 5.2 章节提到，我们的权重设置情况也是在预实验中探索得到。后续研究中可以进一步探讨更合理、简便的惩罚权重的设置方法，我们在讨论部分补充了这一点。

意见 3：第 4.3 章节，作者在使用 MC1 和 MC2 进行参数估计时，其估计方法和收敛情况并未给予呈现。若作者使用 MCMC 算法进行参数估计，此时模型收敛情况将非常重要。因而作者应说明相应的收敛信息。
回应：非常感谢审稿人的建议。MC1 和 MC2 的参数估计采用 MCMC 算法，在 R 中实现参数估计，其 MCMC 设置与 Ozaki(2015)相同，且所有参数估计得到的$\hat{R}$值小于 1.1，达到了收敛标准。已在 4.3 章节补充 MC1 和 MC2 的参数估计方法与收敛信息。

意见 4：第 5 章节，模拟研究二是在题目质量已知的情况设定相应的惩罚参数，现实情境中，题目质量无法预告知晓，此时如何使用惩罚参数？作者应在讨论部分给予说明，以帮助教育工作者更好地使用该方法。
回应：非常感谢审稿人的建议。事实上，在题目质量未知的情景下，惩罚权重的设置需要通过预实验探索得到，使用属性总体掌握程度与总分之间的相关作为权重设置优劣的判断标准，从而调整得到合适的惩罚权重。

例如，在实证研究中，首先根据 MC1 估计得到 15 道题目的失误参数（如下表的第一行数据）将 15 道题划分为 3 个题目质量区，高题目质量：$0 \le \delta < 0.05$；中等题目质量：$0.05 \le \delta < 0.20$，低题目质量：$0.20 \le \delta < 1$。随后将低质量题目的失误权重w_s设置为基准值 1，中等质量题目设置失误权重w_s为$1 + X$（X为正），高质量题目设置失误权重w_s为$1 + X + Y$（Y为正）。由于题目的猜测概率都为$\frac{1}{O}$，所以可以设置所有题目的猜测系数w_g为常数。通过调整 X 和 Y 的值，然后带入$d_{\mathrm{ph-MC}}$的公式进行估计，再计算被试的估计属性掌握程度与总分的相关，当相关达到最高时，此时的失误权重w_s则为最优值。根据此方式，实证研究的失误权重w_s最终被设定如下表第二行。

实证数据失误参数δ以及对应的猜测权重w_s

参数	题目														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
δ	0.97	0.01	0.07	0.11	0.22	0.67	0.03	0.66	0.49	0.38	0.01	0.34	0.57	0.21	0.54
w_s	1	3	2	2	1	1	3	1	1	1	3	1	1	2	1

意见 5：第 7.1 章节，讨论部分过于简略，需对研究中出现的相关结果进行细致分析和讨论，如为何即使在真模型为 MC2 时，MC1 下的诊断结果仍要优于 MC2；以及为何测验长度增加之后，参数估计方法的表现要优于非参数诊断方法；再者还应说明在实际情境中，如何采用作者所提出的方法进行分析等。

回应：非常感谢审稿人的建议。已在第 7.1 章节补充了对研究结果的分析和讨论。关于新方法在实际情景中的适用场景在 7.2 结论部分做了说明。例如，我们指出：在题目质量较差，测验长度较短时，推荐使用简单汉明d_{h-MC}。

意见 6：第 7.2 章节的第 4 个结论中，作者指出带惩罚系数的非参数汉明距离法更拟合这批实证数据；作者做出该结论的依据是什么？实证研究中，作者并没有说明该方法更拟合实证数据。

回应：非常感谢审稿人的提问。我们用估计得到的被试属性掌握程度计算了与被试总分的相关，带惩罚系数的汉明距离得到的相关最高，因此可以判断该方法更“拟合”这批数据。修改稿中，我们对第 7.2 章节的第 4 个结论进行了修改。

意见 7：第 6.2 章节，实证研究中作者使用了惩罚性汉明距离的的分类方法进行测验分析，但惩罚系数的权重如何设定，作者尚未交待清楚。

回应：非常感谢审稿人的提问。请见对意见 4 的回复。

意见 8：部分参考文献的顺序出错。

回应：非常感谢审稿人的提醒。修改稿中已更正。

审稿人 2 意见：本文针对具有干扰选项的多项选择题，基于 Ozaki(2015)的多项选择认知诊断模型提出理想得分的计算方法。在此基础上，将 Chiu 和 Douglas（2013）提出的 3 种基于汉明距离的非参分类方法推广应用到具有干扰选项的多项选择题的认知诊断，具有一定的创新性。在项目质量无差异和有明显差异以及实证题库下比较了参数化和非参数化的方法，实验内容比较丰富。针对本研究，有几个问题和作者进行商讨和学习：

意见 1：多项选择题的反应向量的结果形式是什么？具体而言，每个题目的反应是选项的类别，还是多个 0、1 值？例如，在一个 4 个选项的题目，令 1，2，3，4 代表 4 个选项，反应y_{ij}的值是取对应选项的数值呢，还是y_{ij}是一个四维向量，如果选择 A，那么$y_{ij} = (1,0,0,0)$的形式？文章应该交代清楚。

回应：非常感谢审稿人的提问。多项选择题的反应向量$\boldsymbol{Y}_{ij}$的结果形式是一个多维向量，以 4 个选项的题目为例，$\boldsymbol{Y}_{ij}$是一个四维向量，Y_{ijo}是其中的元素，若选择 A，则$\boldsymbol{Y}_{ij} = (1,0,0,0)$，$Y_{ij1} = 1$，$Y_{ij2\ldots4} = 0$。已在第 3 章中补充了相关信息。

意见 2：论文中关于非参数的分类方法和本文三个指标的写作有重复之处，建议凸出本文的创新点，将写作应突出本研究的核心思想和创新点。

回应：非常感谢审稿人的建议。本文的行文逻辑为：首先介绍传统的非参数方法，进而介绍 MC 测验下的非参数方法，从逻辑上具有递进的关系而非重复。先介绍 Chiu 和 Douglas（2013）在 0-1 计分测验下非参数分类方法有利于读者对传统方法的理解，在此基础上进一步理解本研究的新方法，若直接介绍新方法可能会导致读者思维的跳跃或不理解新方法为什么是这样

的。因此，我们认为先介绍传统的非参数方法再介绍 MC 测验下的非参数方法是有必要的，符合逻辑的。

意见 3：研究设计中将样本量设置为 30，50 和 100 的原因是什么？我的理解是样本量实际上是方法的实验结果的抽样，样本量越小反映的是实验结果的随机误差，而不能体现方法的真实结果。实验的目的是得出方法的真实判准率，那应该使用足够的样本量以减小随机误差的影响。当方法是有效的，那么在实践中应用于小样本诊断也是可行和合理的。特别地，对于模型的诊断方法，更要利用样本数据进行估计，进而得出模型的真实判准率。换句话来讲，我认为小样本会不能反映出参数化和非参数化方法的真实判准率，会降低它们的真实判准率，特别是参数化方法的判准率。当然，这是我的理解，需要于研究者进行沟通。

回应：非常感谢审稿人的专业意见。研究设计中将样本量设置为 30，50 和 100 的原因在于：本研究着眼于提出适用于小样本情境下的诊断方法，以更好的应用于班级水平的教学补救，其实验目的也正是想在小样本前提下，探讨参数类和非参类方法的判准率差异。因此，我们采用了与 Chiu 等（2018）研究中的相同设置。为了避免随机误差的影响，我们在模拟研究中设置了 100 次循环，其目的就是为了尽可能减小随机误差。参数类方法需要足够大的样本量，如饱和模型 GDINA 至少需要几百甚至上千人才能较准确地估计参数，而在小样本情境下，GDINA 模型的表现将会变差。因此，在足够样本量的情况下得到的判准率并不能反映该方法在小样本下的可行性与合理性。

正是由于参数类方法需要较大样本才能较为精准的进行参数估计，在小样本前提下，参数类方法的参数估计精度较低，而非参类方法无需进行参数估计，其判准率基本不受样本量的束缚，因此在小样本情境下，非参类方法更具优势，这也是非参类方法在班级层面（样本量通常不会很大）进行诊断的优势。

意见 4：研究的模型估计方法中是假设题目参数已知的情况下，进行 KS 判断的吗？还是题目参数未知的情况下开展的实验？

回应：非常感谢审稿人的提问。本文提出三种非参数方法均是在假设题目参数未知的条件下进行的 KS 判断，例如，d_{h-MC}距离法所使用到的的公式 4，5，6，7 并不涉及到题目参数信息。

意见 5：研究以中“当题目长度为 10 时，使用 Q 矩阵的后 10 题，当题目长度为 20 时，使用 Q 矩阵的后 20 题，题目长度为 30 时，使用整个 Q 矩阵”的设计理由是什么？

回应：非常感谢审稿人的提问。本文提出的非参数方法的比较对象是 Ozaki（2015）提出的 MC1 和 MC2 诊断模型，因此本文使用了和 Ozaki（2015）研究中相同的 Q 矩阵设置方式和测验长度。

意见 6：研究一中属性掌握状态的设置如下：

被试的 KS 真值$\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$可被定义为：

$$\alpha_{ik} = \begin{cases} 1, & if\theta_{ik} \geq \Phi^{-1}\left(\frac{k}{K+1}\right); \\ 0, & otherwise. \end{cases}$$

为什么是分部函数在$\frac{k}{K+1}$处函数值的倒数？还是属性具有线性层级结构？这样设计的理由是什么？

回应：非常感谢审稿人的提问。本文所采用的被试属性掌握状态的生成方法是基于多元正态阈值模型（Chiu et al., 2009）生成，该方法被广泛应用于认知诊断领域中（e.g. Chiu et al., 2009; Chiu & Douglas, 2013；Chiu et al., 2018；Chang et al., 2019）。

式中$\Phi^{-1}\left(\frac{k}{K+1}\right)$代表$\alpha_{ik}$取 1 时$\theta_{ik}$的临界点，这样设置的目的是使属性间$\alpha_{ik}$取 1 的概率各不相同。例如，若共有 5 个属性，被试在第一个属性$\alpha_{i1}=1$的概率为 1/6，被试在第二个属性$\alpha_{i2}=1$的概率为 2/6，依次类推。这样设置能更真实的反映现实中被试掌握不同属性的概率不同的情景。这样设置并不代表属性间存在线性层级结构，独立结构中也可以存在属性间不同掌握概率的情况。已在第 4.3 章加以补充信息。

意见 7：“所有实验循环 100 次”是同一批人同一批题目循环 100 次，还是人和题都在相同条件下变化重复 100 次？

回应：非常感谢审稿人的提问。“所有实验循环 100 次”是指人和题都在相同条件下变化重复 100 次，并非同一批人同一批题目循环 100 次。

意见 8：第三部分指出“只要被试 *i* 的 KS 与选项 o 所考察的 q 向量不完全一致，或者被试 *i* 随机猜测时，有$\eta^*_{ijo}=0$”当被试随机猜测的时候，那在每个选项上的理想反应都是 0 ？这句话是什么意思？

回应：非常感谢审稿人的提问。第三部分“只要被试 *i* 的 KS 与选项 o 所考察的 q 向量不完全一致，或者被试 *i* 随机猜测时，有$\eta^*_{ijo}=0$”可作如下理解：首先，只要被试 *i* 的 KS 与选项 o 所考察的 q 向量不完全一致，有$\eta^*_{ijo}=0$，例如，题目 *j* 的 q 向量如下所示：

题目 *j* 的 q 向量示例

选项	属性 1	属性 2	属性 3	属性 4	属性 5
A	1	1	0	1	0
B	1	1	0	0	0
C	1	0	0	0	0
D	0	0	0	0	0

A 选项为正确答案。当 KS 为(1, 0, 0, 0, 0)，此时的 KS 与选项 1，选项 2 的所考察的 q 向量不完全一致，则$\eta^*_{ij1}=0$，$\eta^*_{ij2}=0$，$\eta^*_{ij4}=0$。而 KS 与选项 3 所考察的 q 向量完全一致，则$\eta^*_{ij3}=1$。

其次，再例如 KS 为(0, 1, 0, 0, 0)时，与该题目所有选项所考察的 q 向量都不完全一致，则发生随机猜测，所以$\eta^*_{ij1\ldots4}=0$。

意见 9：研究 2 中因题目质量的差异确定的$d_{\mathrm{ph-MC}}$方法的权重，那么非参数的方法事实上运用在信息极其有限的情况下，那么实践中如何确定权重？

回应：非常感谢审稿人的提问。对您该问题的回复可参见审稿人 1 的意见 4。

意见 10：对于某个选项的 q 向量为全 0 的情况如何解释？这时候，在这样选项上是不是不会提供任何诊断信息？

回应：非常感谢审稿人的提问。当某个选项的 q 向量为全 0 时，代表该选项未被编码或是该选项没有考察该题目所需的任何属性，从这个角度来说，该选项不提供诊断信息。

意见 11：在当多级评分项目的每个得分都有一个对应的 q 向量时，本研究的方法本质上与多级评分项目的方法由什么区别呢？

回应：多级评分项目模型处理的作答数据是实际得分数据，如 0 分，1 分，直到满分，得分等级之间有顺序关系，其建模逻辑与 MC 测验诊断方法的建模逻辑不同。多级评分项目模型有三种建模逻辑：例如，基于等级反应模型（graded response models）的 P-DINA（涂冬波等，2010）、基于连续比率模型（continuation ratio models）的序列 GDINA 模型（seq-GDINA, Ma & de la Torre, 2016）、基于分布评分模型（partial-credit models）的 GPCDM（(General Partial Credit Diagnostic Model，高旭亮等，2019）。等级反应模型侧重于分析某个等级及以上所有等级与该等级以下(不包括该等级)所有等级之间的关系，这类模型是从整体出发考虑模型的建构，更适用于分析不强调具体解题步骤的诊断测验，如写作水平测验。连续比率模型侧重于分析某个等级以上(包括该等级)与该等级的向下一个等级之间的关系，分部评分模型侧重于分析两个相邻类别之间的关系，这两类模型都是基于解题步骤(steps)来考虑模型的建构，但连续比率模型更强调作答过程是连续步骤(consecutivesteps)，即只有成功地完成前面的所有步骤，才能成功地执行下一步，它适合分析解题步骤之间具有严格顺序关系的题目，如数学计算题；而分部评分模型是基于一个局部步骤(localstep)来建模，即被试在当前步骤的解答只和前一步有关，这类模型更适合分析相邻步骤之间具有依赖关系的题目，更适合用于分析量表类型的题目。

而 MC 测验的诊断方法处理的作答数据是称名数据（即各个选项），被试对选项的选择是独立的，不存在顺序关系。

第二轮

审稿人 1 意见：

作者经过一轮修改，论文的可读性、流畅性和质量等方面得到较大地提升，作者也较好地回答了审稿人提出的问题。当然，审稿人认为文章还存在进一步提升和修改的地方，具体如下：

意见 1：引言部分，作者将非参数诊断方法分为聚类方法（Chiu, et al., 2009;康春花等，2015）、谱聚类方法（郭磊等，2018）、汉明距离方法（Chiu& Douglas, 2013; Chiu, et al., 2018）。审稿人认为这种分类方法有待商榷，审稿人认为无论是传统的聚类方法（Chiu, et al., 2009; 康春花等，2015）还是新近的谱聚类方法（郭磊等，2018），均可纳入到聚类方法范畴；而 0-1 计分下的汉明距离（或海明距离）和多级计分下的曼哈顿距离（康春花 ,杨亚坤, 曾平飞, 2019）均可纳入到距离判别范畴；目前还有一种非参数诊断方法就是机器学习，如基于 BP 神经网络的认知诊断方法。对此，审稿人觉得将非参数诊断方法大体可归为聚类分析法、距离判别法和机器学习法三种也许是相对可取的分法。这种分类方法仅供作者参考。

康春花, 杨亚坤, 曾平飞. (2019). 一种混合计分的非参数认知诊断方法: 曼哈顿距离判别法. *心理科学*, *42*(2), 455-462

回应：非常感谢审稿人的建议。我们已在第 1 章节和第 2 章节中修改了相关的表述，并添加了相应的引文。

意见 2：审稿人发现公式（4）-（6）稍微有些复杂，可直接对公式（6）进行简化而得到以下两个公式：

$$\eta_{ijo}^{*} = \eta_{ijo} * \left[1 - \prod_{k=1}^{K_j} (1 - \alpha_{ik})\right] \qquad (A)$$

$$\eta_{ijo} = \prod_{k=1}^{K_j} [2 - 2^{(\alpha_{ik} - q_{iko})^2}] \quad \text{（B）}$$

回应：非常感谢审稿人的仔细审核和改进建议。我们仔细推导了审稿人给出的简约公式，其本质和原稿中的公式（4）-（6）相同，因此我们将原稿中的公式（4）-（6）按照审稿人的公式进行了修改，并对式中参数的描述做出了调整。此外，由于在第 3.3 章节的公式（8）以及第 4.3 章节的公式（9）和公式（10）中涉及到第一轮修改稿公式（5）的γ_{ij}参数，所以我们在第 3.3 章节的公式（8）下方补充了对γ_{ij}参数的说明。

意见 3：公式（5）和（6）中，连乘符号使用错误。举例而言，假设两个题目是 j1 和 j2，当$\boldsymbol{q}_{j1,o} = (1,0,1)$和$\boldsymbol{q}_{j2,o} = (0,1,1)$时，两个题目的$K_j$均等于 2，此时题目 j1 的 k 值应该等于 1 和 3；而题目 j2 的 k 值应该等于 2 和 3；但按照作者的写法，则两个题目的 k 值均为 1 和 2，这是不对的。

回应：非常感谢审稿人的指正。我们采用K_j^*代替了之前的K_j，并在第 3 章节中对该参数的含义进行了修改，沿用了 de la Torre（2011）的思路：K_j^*表示题目 j 所考察的属性个数，有$K_j^* \leq K$，即将元素 0 去掉并将元素 1 向前排序，以使这些考察了的属性为前K_j^*个属性（de la Torre，2011）。并通过举例的方式，对 k 的取值进行了说明。

意见 4：公式（6）下面第二行，作者说“η_{ijo}用于判断被试 i 的 KS 与选项 o 所考察的 q 向量是否完全一致”，这句话也是错误的。应该说是被试 i 的属性掌握模式在 collapsed 之后的属性掌握状态（记为AMP_j^*）与题目所考察属性之间是否完全一致。

回应：非常感谢审稿人的指正。已在第三章节中，对η_{ijo}的表述进行了更改。

意见 5：实验设计部分（4.3 节），作者在生成被试真实的属性掌握模式时，使用了多元正态阈值模型，其中使用了函数$\Phi^{-1}(.)$，作者应该在正文中对这个函数进行阐述和说明。否则，读者可能不清楚这个函数所表达的意思。

回应：非常感谢审稿人的建议。已在第 4.3 章节中补充了$\Phi^{-1}(.)$函数的含义。

意见 6：在表 3 和表 4 所呈现的结果中有相同的趋势：当题目质量较高和测验长度较长时，真模型得到的 PCCR 表现在四种诊断方法中表现最好，而在其它所有条件下，d_{h-mc}方法的诊断结果最好。对于该结果，审稿人认为作者可在讨论部分对其原因或意义进行讨论。

回应：非常感谢审稿人的建议。已在第 7.1 章节的第一段对该问题进行了讨论。

意见 7：实证研究部分应对惩罚汉明距离的具体操作方法进行相应说明，以使读者明晰该方法是如何应用于实证数据的。

回应：非常感谢审稿人的建议。在第一轮审稿的审稿人 1 第 4 个意见的回复中，我们详细阐述了惩罚汉明距离方法在针对这批实证数据的惩罚权重设置方法。具体而言，我们将实证数据的 15 道题划分为 3 个题目质量区，高题目质量：$0 \leq \delta < 0.05$；中等题目质量：$0.05 \leq \delta < 0.20$，低题目质量：$0.20 \leq \delta < 1$，并对三个区间的题目设置不同的惩罚权重。具体操作步骤已补充在 7.1 章节第三段。同时，我们指出，若更换了测验，需要按照操作步骤调整 X 和 Y 的取值，以使惩罚汉明距离方法在新的测验中达到最优表现。

意见 8：7.1 节最后一段，作者指出“CD-CAT 需要大样本才能发挥其优势，小样本下的表现较差”。审稿人并不赞同作者的这个观点。审稿人认为，当题库中的题目参数得到较为准确地校准的情况下，即使正式测验的样本量很小，参数 CD-CAT 也可以很好地对被试进行诊

断分类，而且审稿人觉得此时的分类结果要优于非参数 CD-CAT 的结果。只有当题库中各题目参数的校准误差比较大时，参数 CD-CAT 的分类结果可能比非参数 CD-CAT 的结果更差。如 Chang 等（2019）使用 30 个被试作答数据和 GDINA 模型对题库中的 300 道题目的参数进行校准，此时可以预期其校准误差会比较大。此刻其非参数 CD-CAT 的分类结果才优于参数 CD-CAT 的分类结果。

回应：非常感谢审稿人的指正。已对第 7.1 节最后一段进行修改。

意见 9：部分语句的表达需提炼，如引言部分第三段“主流题型为选择题(Multiple-Choice,MC)题型”可写为“主流题型为选择题(Multiple-Choice,MC)”；公式（2）下面第一行“则该题目对区别出答对和答错的被试的能力就越强，”可写成“则该题目对答对和答错被试的区分能力就越强”；公式（8）下面第二行同样存在这种问题。

回应：非常感谢审稿人的建议。已对上述语句表达进行了修改。

意见 10：部分参考文献的顺序依旧存在错误，如 Li, Y. (2014)应该置于 Liu, T. (2016)的前面；又如 Chiu, C.-Y., Douglas, J. A., & Li, X. D. (2009).应置于 Chiu, C. Y., Sun, Y., & Bian, Y. (2018).的前面。

回应：非常感谢审稿人的指正。已对参考文献顺序进行了修改。

审稿人 2 意见：本文针对多项选择题中具有干扰项的分析提出了非参数诊断方法，通过两个模拟研究不仅比较了非参数方法和参数方法的优劣，还检验了非参数方法在不同条件下的稳定性，最后通过实证数据再次验证非参数方法的表现。作者对审稿意见做了详细的分析和答复，并在文中进行了恰当的修改。总体上讲，研究讲基于汉明距离的非参数方法推广应用至具有干扰选项的多选题分析，研究具有实际应用价值和意义，实验设计合理，行文符合逻辑，可读性较强。同意发表。

回复：非常感谢审稿人的认可。