《心理学报》审稿意见与作者回应

题目：认知诊断测验中的项目功能差异检测方法比较

作者：王卓然 郭磊 边玉芳

第一轮

审稿人 1 意见：

意见 1：在 P10 3.2.3 部分，作者考察了四种存在 DIF 的实验条件，但为什么未考虑 s 减少 g 增加的情形？该条件是存在一致性 DIF 另一种条件，本文对一致性 DIF 的研究结果是否能推广到该条件下呢？

回应：谢谢专家提出的问题。在认知诊断情境下构造 DIF，一般都是先生成对照组的题目参数，再在对照组题目参数上经过一定的变化生成目标组的题目参数。在这种模式下可以构造 8 类 DIF，分别是，s 增加，g 增加，s 增加 g 增加，s 增加 g 减小，s 减少，g 减少，s 减少 g 减少，s 减少 g 增加。由于这些 DIF 构造情景都是在相对于对照组改变目标组的题目参数，所以当目标组和对照组位置对调后，后四类情景就跟对调前的前四类情景相同了。例如，s 增加即为目标组 s 比对照组 s 大 DIF 量，也就是对照组 s 比目标组 s 小 DIF 量，相当于基于目标组对对照组做 s 减少的变化。在模拟研究中，目标组和对照组的位置是对等的，特别是在本研究中目标组和对照组的受测者数也相同，所以我们认为目标组和对照组是可以互换的。因此，在本研究中，我们只讨论了 s 增加，g 增加，s 增加 g 增加，s 增加 g 减小这四类 DIF 的构造方法，而没有讨论与之对应的另外四类 DIF 构造方法。Zhang(2006)也是使用了这四类 DIF 的构造方法，相信他也是基于以上原因。根据专家的问题，在文章的 3.2.3 部分，我们新增了对于这种省略的说明。因此对于一致性 DIF 的研究结果也可以推广到 s 减少 g 增加的条件下。

意见 2：在 P10 3.2.4 部分，作者做 DIF 检测时，各方法的参数估计的数据是按参照组和对照组两组分开估计，还是一起估计？如果参数估计是按参照组和对照组两组分开估计的话，你是否进行了等值？如何等值的？

回应：谢谢专家提出的问题。本研究中采用了 3 种 DIF 检测方法，其中 MH 法和 LR 法都是把对照组和目标组的数据一起估计。而 Wald 检验法对于目标组和对照组是分开估计的，并且不需要等值。Wald 检验法的原理是，分别估计出目标组和对照组的两套题目参数，然后检验这两组参数之间差异是否显著。

意见 3：P10 3.2.4 部分，LR 法具体是如何操作的呢？如果是采用回归，作者是对原始数据进行回归，还是基于匹配变量分层的结果进行回归？如果是前者，那么匹配变量对 LR 法不存在影响，也就不需要考虑匹配变量了；如果是后者，那么回归所使用的数据是如何获取的，能具体解说一下吗？

回应：谢谢专家提出的问题。LR 法是分两步进行的：第一步是计算得出匹配变量，第二步是把匹配变量作为预测变量被试作答作为结果变量进行回归分析。使用 LR 法时，匹配变量是加入到回归方程中的变量，而在使用 MH 法时，匹配变量被用来给受测者分层。在本研究中匹配变量有总分，能力值θ，和知识状态 KS 三种。总分通过加和受测者在每道题目上的得分就可以得到，使用 2PL 模型估计出能力值θ，使用 RSM、AHM 和 DINA 模型三种方法估计出知识状态 KS。

意见 4： 4.3LR 法对于一致性 DIF 和非一致性 DIF 检测的检验力和一类错误率，是结果的重复表述。4.3 中部分结果表述与 4.1 不一致。(从审稿专家在原文中做的批注中提炼)
回应：感谢专家提出的问题。由于 MH 法和 Wald 检验都不能区分一致性 DIF 和非一致性 DIF，为了便于三种方法的比较，4.1 和 4.2 展示的是三种方法目的是检测出 DIF 时的表现。为了展现出 LR 法可以分别检测出一致性 DIF 和非一致性 DIF 的特性，在 4.3 中单独对于 LR 法对于一致性 DIF 和非一致性 DIF 检测进行了展示。正是由于 4.1 和 4.2 关注的是检测出 DIF 的能力，而 4.3 中关注的是 LR 法对于一致性 DIF 和非一致性 DIF 的区分能力，4.1、4.2 和 4.3 中部分规律才会有一定的差别。

审稿人 2 意见：
意见 1：本文通过修改 logistic regression method (Swaminathan & Rogers, 1990)来实现对 CDA 中项目的 DIF 侦查，想法很直接。审稿人认为，LR 法可适用于大多数 IRMs (如作者所述的“不基于模型的 DIF 检测方法”)也是因为绝大多数 IRMs 都属于 logistic regression model。作者直接将 LR 法应用于 CDMs 就存在一个前提假设“本研究承认 person attribute profile 对项目反应概率的影响满足 logistic regression”。但从当前稿件看作者或许并未认识到这点，因为作者将 LR 法也应用到了 RSM 和 AHM 等非 logistic regression 的 model 上。尽管审稿人没有重复作者的模拟研究，但从作者给出的结果看，将 LR 法应用于 RSM 和 AHM 所得到的结果尚可。那这种结果是否寓意了 CDMs 其实质就是一种简化了的多维 IRMs 呢？这有待作者进一步思考。
回应：感谢专家的点评。我们认为由于 LR 法是不基于模型的 DIF 检测法，所以使用 LR 法进行 DIF 检测的可行性与拟合受测者作答数据所采用的模型和方法没有关系。使用 LR 法进行 DIF 侦查是基于 DIF 最基本的定义，即来自不同群体但能力水平相同的受测者，在同一题目上具有不同正确作答概率。LR 法检测 DIF 时通过把题目作答作为结果变量，把受测者能力和所属的群体作为预测变量，进行回归分析。如果受测者所属群体对应的回归系数显著不为零，就说明受测者在这道题目上作答的正确概率在控制了能力的情况下，还受到所处群体的影响，也就是存在 DIF。使用不同的模型或方法(例如 IRMs、DINA、RSM 和 AHM)，只是为了估计出受测者的能力。就像专家提到的一样，认知诊断相当于 IRT 离散化多维化的拓展，从本质上来讲仍然是对于受测者能力的分析。因此，person attribute profile（KS）对于项目反应概率的影响与 IRT 中的能力值θ一样，也应该满足 logistic regression。虽然我们同意专家的意见，认为 CDMs 可以看做是一种简化了的多维 IRMs，但是这个推论与本研究的结果没有关系。

意见 2：本文写作水平较低，排版、格式、公式等出现较多问题。建议作者今后提高稿件的写作水平，以便审稿人更为清晰地理解您所要表达的意思且不影响审稿人阅读稿件时的情绪；
回应：感谢专家提出的意见。我们已经认真修改过排版、格式、公式等问题。对于之前疏漏给专家带来的不便，我们深表歉意。

意见 3：该稿件处于未完成状态，如“3.2 数据模拟”处缺少详细内容；
回应：感谢专家提出的问题。可能是由于标点符号使用的问题，使专家产生了误解。文章的结构是完整的。3.2 数据模拟包括：3.2.1 测验 Q 矩阵、3.2.2 受测者 KS 真值模拟、3.2.3DIF 及受测者作答模拟、3.2.4DIF 检测。也就是说，3.2 部分将测验 Q 矩阵的构建，受测者 KS 真值的模拟，DIF 和受测者作答的模拟，以及 DIF 检测的整个模拟研究过程都描述了出来。

第二轮

审稿人 1 意见

谢谢你对专家审稿意见的详细回答，该研究有一定的创新性，其研究结果有一定的参考价值。

审稿人 2 意见

意见 1：Hou, de la Torra, Nandakumar(2014)一文中探讨了 Wald 对一致和非一致性 DIF 的处理。

回应：感谢专家提出的问题。此文中确实提到 Wald 检验法可以检测出一致性 DIF 和非一致性 DIF。但 LR 相比于 Wald 的优势在于 LR 法可以区分出一致性 DIF 和非一致性 DIF。所谓区分，指的是能够通过统计检验的方法得出是一致性 DIF 还是非一致性 DIF 的结论。举个例子，题目 1 有一致性 DIF，题目 2 有非一致性 DIF。采用 Wald 检验，可以给出题目 1 和题目 2 都含有 DIF 的结果，但无法给出题目 1 是一致性 DIF 还是非一致性 DIF 的结果。而采用我们引进的 LR，可以给出题目 1 含有一致性 DIF，题目 2 含有非一致性 DIF 的结果。所以我们说 LR 具有区分一致性 DIF 和非一致性 DIF 的功能。因为一致性 DIF 和非一致性 DIF 产生的原因和带来的影响都是不同的，所以我们认为区分一致性 DIF 和非一致性 DIF 还是很有意义的。

意见 2：表 1 制作仍不符合论文发表规定

回应：感谢专家的意见。我们已将表 1 改为三线表。

意见 3：论文书写的逻辑性存在问题。正常应该是先介绍已有方法，然后在单独一章列出已有方法的局限性和进而提出的新方法。建议单独列出，以体现本文提出的新方法。

回应：感谢专家提出的问题和建议。在引言中我们已经介绍了已有的方法以及其不足，并提出了新方法的优势。第二部分介绍 DIF 检测方法只是为了给出本研究中使用的方法的具体说明，便于读者理解。因此将已有方法和新方法一并列出。我们也考虑过将新方法放到研究设计部分进行介绍，但总感觉不如新旧方法比对介绍来的更加直观，也便于读者了解这些方法。

意见 4：审稿人发现作者采用 Wald 法得出的结果比 Hou et al., (2014). Differential Item Functioning Assessment in Cognitive Diagnostic Modeling: Application of the Wald Test to Investigate DIF in the DINA Model. 一文中的高，即作者说的“过度膨胀”。请问，作者是否采用了 CDM package？该包对 SE 计算存在问题，会低估 SE，导致 Type1 error 增大。

回应：感谢专家提出的问题。Hou et al., (2014)列出的结果是分不同的题目参数水平、题目考查属性个数和被试人数而不同的，而本研究中的题目参数设定为 0.1~0.3，题目中 5 道一题一属性、10 道一题两属性、10 道一题三属性。因此，按照本研究的条件，Hou et al.的结果中一类错误率应该为.068~.145 左右，仍然高于 LR 法得到的.03~.10 的水平，仍然是过度膨胀的结果。我们确实采用了 CDM package。CDM package 是根据 de la Torre (2008)的 EM 算法而编制的。我们没有找到有关 CDM package 或 de la Torre (2008)的 EM 算法 SE 估计有偏的文章，麻烦专家给出相关文献信息，便于我们学习。

意见 5：LR 的局限之一在于只能进行“检测”，无法报告 DIF 的 effect size。建议作者在对 LR 法修改后，应给出该方法可能存在的不足，以便于后期测验分析人员使用。

回应：感谢专家提出的问题。LR 确实存在不能报告 effect size 的问题。但是，LR 有像 MH 法和 SIBTEST 方法一样关于可忽略的 DIF、中等的 DIF 和较大的 DIF 的划分（Hidalgo, M. D.，

2004，Zumbo & Thomas, 1997）。相信这种划分也可以提供有关 DIF 严重程度的信息。在研究方向展望部分，我们加入了对于 LR 可能存在不足的讨论。