《心理学报》审稿意见与作者回应

题目：认知诊断评价中测验结构的优化设计
作者：彭亚风;罗照盛;喻晓锋;高椿雷;李喻骏

第一轮

审稿人 1 意见：
意见 1：在测验编制上有一定实践价值；
回应：感谢审稿专家的肯定。

意见 2：国内学者如涂东波、蔡艳、颜远海、丁树良等人在这个方面有一定的研究进展，文献综述中没有体现出来；
回应：已补充，并增加引述了 Liu, Huggins-Manley 和 Bradshaw 新近发表的文章（详见前言的第 5-6 段蓝色字体部分）。

意见 3：文中关于不同属性层级模型的 PMR 的差异和之前研究结论不一致，DINA 模型本身不考虑属性间层级关系，应对此“迟钝”；
回应：颜远海，丁树良和汪文义（2011）中指出 DINA 模型对属性层级关系“迟钝”，具体表现在计算了每个属性的判准率后，发现当属性层级结构紧密的情况下，父节点的判准率会低于子节点的判准率。但是本研究是以模式判准率为评价指标，没有细致到每个属性上的判准。并且在模式判准率上的研究结论与蔡艳, 涂冬波和丁树良（2013）以及颜远海等人（2011）的研究结果一致。

意见 4：由于认知诊断测验多出现在单元测试中，在 Q 中加入多个 R，意味着在这种单元测试中，重复出现多组类似题目，在测验编制实践中，这种做法并不常见；
回应：本研究的目的是从理论比较六种属性层级关系下的不同测验设计方案，找出针对每种属性层级关系的测验结构设计的优化方案，为实际的测验编制提供一般性的设计准则。在实际的测验编制情境中，考虑测验编制的难易程度，可以藉由当前的研究结果为编制者选择合适的测验编制方案提供较为妥当的参考意见。

意见 5：文中提到的奇数定律，一般的表述是，只要 Q 中有 R，题目并不是越多越好。奇数定律如果成立，可能的理论解释是什么。
回应：得到奇数规律的结果后，我们对奇数规律的成立进行了讨论，在对被试作答数据进行分析时，发现当测验 Q 矩阵仅由偶数个 R*构成时，R*中的每种项目考核模式会出现偶数次，会有被试在某种考核模式上做对一半做错一半，这类被试在 AMP 被判错的被试群体中所占比例较高。由此推论奇数规律成立的可能的原因是：以 5 个属性独立型 20 个项目为例，当测验仅有 4 个 R*构成时，项目考核模式为[1 0 0 0 0]会出现四次，假设这 4 个项目对应的项目参数分别为：s_j和$g_j\left(j = 1\cdots 4\right)$。对于掌握和未掌握该项目考核模式所考察的属性的被试来说，被试在这 4 个项目上作答的似然函数分别为：$L_1 = \prod_{j=1}^{4}\left(1 - s_j\right)^{x_j} s_j^{1-x_j}$、$L_0 = \prod_{j=1}^{4} g_j^{x_j}\left(1 - g_j\right)^{1-x_j}$

被试在这 4 个项目上的作答有以下几种情况：答对 4 题、答对 3 题答错 1 题、答对 2 题答错 2 题、答对 1 题答错 3 题以及答错 4 题。当出现答对两题和打错两题（假定答对前两题）时，$L_1 = (1-s_1)(1-s_2)s_3s_4$，$L_0 = g_1g_2(1-g_3)(1-g_4)$。在项目质量比较均匀的情况下，此时 L_1 与 L_0 之间的差异很小，无法区分出被试的作答是基于掌握了该属性还是未掌握，而测验中又没有其他考察第 1 个属性的项目，这就无法为被试在第 1 个属性上的诊断提供信息。当测验包含最大奇数个 R*时，并不会出现上述作答情况。以上是关于奇数规律成立的可能的解释，后续需要对奇数规律成立的原因进行系统的研究。

审稿人 2 意见：本研究讨论 Q 矩阵中测验长度、类 R 阵的个数以及类 R 阵外项目的属性个数对认知诊断测验效果的影响，研究的问题有价值，研究方法也比较妥当。但是，总的来看，研究的设计还可以更加丰富、写作上也存在不少问题。具体意见如下：

意见 1：对于 Q 矩阵以及可达矩阵的相关研究搜集和回顾的不够，特别是国内有关学者的工作，不宜忽略。

回应：已补充，并增补了 1 篇 2016 年发表的外文文献。（详见前言的第 5-6 段蓝色字体部分）

意见 2：该研究以 PMR 均值和标准差作为诊断效果的唯一评价，需要对 PMR 参数给予足够的介绍、说明；作者在说明时又给出了 APM 值，却没有再对 APM 给予任何解释。

回应：PMR 是指所有属性都判准的被试比例，PMR 越大，表明分类准确性越高。已在文中对 PMR 进行解释。非常抱歉，APM 为笔误，应该为 AMP，已更正。

意见 3：当前的实验设计中每个测验的类 R 阵为单一一种属性层级关系，但实际应用中可能会多种并存，是否可以在仿真实验中也考虑并存的情况。

回应：经审稿专家提醒，在模拟实验的部分增加了一种多种属性层级关系并存的“混合型”属性层级关系，并于文末附上了对应的属性层级关系示意图。

意见 4：文章的整体写作更像是个实验报告，而非适合于《心理学报》发表的论文；没有必要在每一个小节开始处都写研究目的（实验目的）；特别是讨论部分缺少与已有研究成果的比较。

回应：已做对应的修改。

意见 5：在当前的“讨论”一节中，声称发现了分类准确性呈现奇数规律，但仅在如此有限的仿真结果下得此结论是否过于武断；同时对结果的分析和讨论，还忽视了不同层级关系结果的比较。

回应：为了提高实验结论的说服力，我们采纳审稿专家的意见，对实验条件进行了扩充，增加了属性个数、项目质量这两个自变量，并设定了不同的水平，以模拟更加复杂的测验情景。结果显示，奇数规律仍能成立。在结果的分析和讨论中，增加了不同层级关系结果的比较。

意见 6：文中的图 1、图 2、图 4 均使用颜色区分不同的测验长度、层级关系，但使用黑白打印机打出后，很难分辨清楚。

回应：已修改

意见 7：英文摘要部分的问题：一是出现了对文献的引用，一般而言摘要是自己研究的目的、

过程和结果，没有引用别人的原话的必要；二是存在很多中式英语，如论文接收，可考虑请专业翻译公司润色。
回应：已修改

意见 8：文中还有一些次要的问题，如对于半角和全角的括号没有有效区分，大量混用；Chiu, Douglas & LI 的引用时，Li 的 I 应为小写；第 1 节中说到 Chiu、Madison 的研究时，未使用正确的引用格式
回应：已修改

第二轮

审稿人 1 意见：
意见 1：基本上已经回答上次所提问题。所探讨内容对属性诊断测验的编制有实际意义，方法适当，结论基本可靠，建议发表。
回应：感谢审稿专家的肯定与支持。

审稿人 2 意见：本文作者对论文进行了认真的修改，特别是增加了实验条件，使得论文的质量有了很大程度的提升，基本达到了发表的要求。还希望作者能做出以下修改：
意见 1：文章摘要部分声称“根据实验结果，本研究提出了进行诊断评价时 Q 矩阵优化设计的一些基本准则”，但在讨论等部分并没有看见明显的准则内容；另外，通过一篇实验研究提出准则是否言语表达的过于强烈，改为提出建议更为合适。
回应：已根据审稿专家的意见进行相应的修改。在讨论部分阐述了在不同属性层级关系下，测验 Q 矩阵优化设计的一些建议。

意见 2：尽可能减少口语化的表达，而使用学术语言，特别是在讨论部分；另外，仔细检查文中的错别字。
回应：我们对全文进行了反复通读，尽可能的使用学术语言进行表达，并认真的检查了文章的错别字。对于存在的错别字，作者为自己的粗心表示歉意。

意见 3：英文摘要还可以进一步扩充，将讨论部分的关键内容翻译为英文，放在其中，对于国外学者了解我们的研究很有价值。
回应：已根据审稿专家的意见，将讨论部分的关键内容翻译成英文放入英文摘要中。