《心理学报》审稿意见与作者回应

题目：多维题组反应模型：多维随机系数多项 Logistic 模型的应用拓展
作者：魏丹 刘红云 张丹慧

第一轮

审稿人 1 意见：
意见 1：从模型本身讲，新模型是对多维随机系数多项 Logistic 模型的约束，并非拓广；新模型是在适用情境上对多维随机系数多项 Logistic 模型做了拓广。因此题目需要修改。
回应：

感谢审稿专家的意见。本文题目表述确实不够准确，两位审稿专家都指出了这一问题。本文核心目标是构建一个多维题组反应模型，该模型是对多维随机系数多项 Logistic 模型在应用情境上拓展，结合两位专家的意见，本文题目改为：多维题组反应模型：多维随机系数多项 Logistic 模型的拓展。文章中中文题目和英文题目都已经做出相应的修改。

意见 2：对于多级评分题目，文中只报告了“项目难度均值”等。多级评分题目与二级评分题目的“难度”是有差异的，所以这里的只报告一个“均值”是什么均值？
回应：

感谢审稿专家的意见。原文中对项目难度呈现结果没有给出清楚的说明。本次修改对两个模拟研究项目参数估计结果的呈现方式和呈现顺序都进行了调整，并进行了更加清晰的描述。研究一中，对二级评分项目计算所有项目难度估计的 RMSE，对多级评分项目计算所有步骤难度的 RMSE。为了更清晰的呈现本文 MTRM 在不同测验形式下对项目参数的估计的差异，对各个测验情境中所有项目参数估计 RMSE 求均值并画图，例如对 0.3b, 0.6b, 0.9b 三个二级评分的情境分别求三个测验中所有项目难度估计的 RMSE 均值，对 0.6p 的情境求该测验中所有步骤难度估计的 RMSE 均值。修改内容见第 9 页 3.1.4 节第一段红色字体内容。研究二修改了参数估计结果的呈现方式和呈现顺序，已经不存在该问题。应用研究中项目参数结构图 9（d）也进行了修改，图中数值由原来的“项目难度均值”改为步骤难度结果，具体见图 9（d）以及相应的描述的修改见 3.3.2 节中红色字体内容。

意见 3：针对目标能力多维度，詹沛达等(2015)在认知诊断框架下提出了多维题组效应认知诊断模型，作者需要在文中提及并说明差异。另外，作者在引言中的部分内容与詹沛达等(2013)的文章相似，需引用。

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

詹沛达, 王文中, 王立君. (2013). 项目反应理论新进展之题组反应理论. 心理科学进展.
回应：

感谢审稿专家的意见。已经对詹沛达等(2013,2015)文章进行了仔细阅读。对原文前言 1.1 节第一、二两段内容进行了删减和合并，并间接引用了詹沛达等（2013）文章中内容，列入参考文献，修改内容见第 1 页 1.1 节第一、二段中红色字体。另外，确实多维题组效应认知诊断模型模型也可以解决目标能力多维的问题，和本文 MTRM 对测验结构的适用范围是相同的，不同的是多维题组效应认知诊断模型是对被试认知属性的掌握情况进行二分或者

多分的判定，而 MTRM 是对被试能力进行连续的估计。因此本次修改在文献综述部分添加了对多维题组效应认知诊断模型的简单介绍，并引用相应文章列入参考文献。修改内容见第 2 页下方红色字体。参考文献如下：

Zhan, P. D., Wang, W.-C., & Wang, L.-J. (2013). Testlet Response Theory: An Introduction and New Developments. *Advances in Psychological Science, 21*(12), 2265-2280.

[詹沛达，王文中，王立君. (2012). 项目反应理论新进展之题组反应理论[J]. *心理科学进展, 21*(12), 2265-2280.]

Zhan, P. D., Li, X.-M., Wang, W.-C, Bian, Y. F., & Wang, L.-J. (2015). The Multidimensional Testlet-Effect Cognitive Diagnostic Models. *Acta Psychologica Sinica, 47*(5), 689-701.

[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, 47*(5), 689-701.]

……

审稿人 2 意见：

意见 1：论文的标题值得商榷，本文的核心目标不是对多项 logistic 进行拓展，而是构建一个多维题组反应模型。所以标题是否可以改为“多维题组反应模型：一种对多维随机系数多项 Logistic 模型的拓展”。这仅仅提供了另外思考问题的角度，请作者自行斟酌。

回应：

感谢审稿专家的意见。本文题目表述确实不够准确，因此两位审稿专家都指出了这一问题。本文核心目标是构建一个多维题组反应模型，该模型是对多维随机系数多项 Logistic 模型在应用情境上拓展，结合两位专家的意见，本文题目改为：多维题组反应模型：多维随机系数多项 Logistic 模型的拓展。文章中英文题目都已经做出相应的修改。

意见 2：本文号称新的模型同时解决多维能力和多维题组的问题，但模拟数据只用了能力项目间多维分析，没有涉及能力项目内多维的内容。

回应：

谢谢审稿专家的意见。本文新模型 MTRM 在能力维度和题组维度上都没有进行多维结构的限定，因此确实可以适用于多维能力和多维题组同时存在的情况，包括能力的项目内多维。确实如您所说，本文模拟数据只用了能力项目间多维分析，没有涉及能力项目内多维。本文模拟研究暂且不考虑项目内多维能力。

一方面，因为本文研究问题的重点在于将题组反应模型中的单维能力拓展为多维能力，主要关注能力项目间多维，这类测验分析在 IRT 框架下的题组反应模型研究领域已经是一个有待解决的问题，因此模拟研究也以能力项目间多维结构为主。

另一方面，从实验设计上来看，因为项目内多维能力之间势必存在更加复杂的、难以控制的相关关系，而本文研究一表明能力维度间相关系数是影响结果一个重要因素，为避免这种复杂的相关关系在对研究结果进行比较分析时带来的混乱，因此在实验设计中没有考虑项目内多维能力，而是关注如何更好的控制无关变量，从而得到影响模型参数估计的因素。

因此，本文模型可以适用于项目内多维能力、高阶等测验结构，这是由于本文模型结构的灵活性，可看作模型的广泛适用性。

意见 3：同样，实际对项目内多维题组的实验也仅限于研究二中的结构 2，其他模型均为项目间多维题组（是蔡力的 Two-tier 模型已解决的问题）；特别是使用的实际数据并不存在项目内多维题组的情况。

回应：

感谢审稿专家的意见。本文确实只有研究二中的结构 2 存在项目内多维题组，暂且不考

虑对模拟研究的模拟测验结构进行改变。因为目前两个模拟研究的这种设计能够更好的控制无关变量，从而得到所关心的影响因素对结果的影响，达到既定的研究目的。研究一和研究二是相辅相成的，研究一中多级评分情境的设计也更好建立了研究一和研究二之间的联系。在研究二证明结构复杂性对结果影响不大的基础上，研究一的结论对于研究二的结构 2 同样成立。另外，因为目前题组研究中涉及到多维能力的都是因子分析框架下的双因子模型和 two-tier 模型，因子分析框架下的结果与 IRT 模型结果在本质上是不同的，而 IRT 框架下的题组反应模型主要关注的还是单维能力的测验，因此将题组反应模型领域的单维能力拓展到多维能力是作为本文主要创新点的。

确实如您所说，本文模型在结构上和 two-tier 模型解决的是相同类型的测验分析，虽然在蔡力（2010）的文章中明确限定了 two-tier 模型次要维度（specific dimension）之间不能交叉，但其应用软件（如 flexMIRT）进行拓展之后同样可以在 two-tier 模型的框架下分析项目内多维题组。因此本文 MTRM 和 two-tier 模型在结构上适用于同类测验。但是，两个模型对于结果的解释上存在本质的差异，从不同的角度对测验进行分析。第一，two-tier 模型更加倾向于从因子分析的角度分析测验，模型中主要维度（primary dimension）和次要维度（specific dimension）都是模型所关注的重点，因此对于次要维度的正交性假设是非常强的假设，会增大模型误差，而在题组反应模型中，题组效应看作能力维度的干扰因素，不关注个体层面的题组效应大小，因此对题组维度的正交性假设也更容易接受；第二，two-tier 在项目参数估计中得到的是项目各个评分等级 k 上的截距参数 c_k 和各个主要维度 p（primary dimension）上的斜率参数 a_p,项目难度参数通过计算$\mathrm{b} = -c/a$得到（Houts & Cai, 2016），这里的难度参数 b 与 IRT 中的项目参数的解释不同，估计值也存在较大差异。如果测验中存在项目内多维能力，那么一个项目对应多个斜率参数a_p，无法直接得到项目难度参数。而本文 MTRM 则从 IRT 分析角度出发，同时得到 IRT 中相同定义的项目参数和被试能力估计结果。

结合您给出的意见 6，为了在文章中更好的体现出两个模型的区别和各自的特点，在本文模拟研究二中加入 two-tier 模型估计结果进行对比。具体添加和修改内容见文章 P11 下-P14 上中红色字体的内容。

另外，确实本文使用的实际数据并不存在项目内多维题组，由于我们使用的数学测验数据不构成项目内多维题组，所以选择项目间多维题组结构进行分析。但是在英语测验或者语文测验中很容易构成项目内多维题组，题组效应的影响也更加明显，对于这些存在项目内多维题组甚至更加复杂的测验结构，本文模型也同样适用并得到很好的估计结果，题组效应更明显时也更加能够体现本文模型的优势。

Li Cai. (2010). A TWO-TIER FULL-INFORMATION ITEM FACTOR ANALYSIS MODEL WITH APPLICATIONS. *Psychometrika*, *75*(4), 581-612.

Houts, C. R., & Cai, L. (2016). flexMIRT user's manual version 3.5: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

意见 4：文中的 conquest 或 Conquest，如果指 IRT 软件的名称，准确写法应该是 ConQuest。另外，应对这个软件有一两句描述，其并不是很多人熟悉的软件。另外，没有看到对模拟数据生成程序的介绍。

回应：

感谢审稿专家的意见。原文对软件名称的写法不够标准，也没有相关的介绍和引用，因此本次修改对文中多处 ConQuest 软件的写法进行了统一和标准化，并引用了该软件的使用手册（如文中 P3 上面第一段中红色字体内容）。另外添加的 flexMIRT 也引用了软件使用

手册，有助于不熟悉这些软件的人查阅相关资料进行学习。参考文献如下。

Wu, M. L., Adams, R. J., Wilson, M., & Heldane, S. A. (2007). ACER ConQuest: Generalized item response modeling software (version 2.0) [computer software]. Melbourne: Australian Council for Educational Research.

Houts, C. R., & Cai, L. (2016). flexMIRT user's manual version 3.5: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

原文对模型数据生成方法进行了简要说明，但是不够具体。此次修改对模拟数据生成步骤进行了更细致和详细的描述，增加 3.1.2 节被试反应数据模拟。具体修改内容见 P8 中 3.1.2 红色字体的内容。本文两个模拟研究中的模拟参数设定有所不同，因此对模拟数据的具体设定的描述在两个"模拟研究设计"部分，在本节中不再进行描述。

意见 5：研究一对结果的呈现顺序为先项目参数、后能力参数，但到了研究二中顺序反了过来，不利于读者理解。

回应：

感谢审稿专家的意见。原文两个模拟研究结果呈现方式不一致，为方便读者，本次修改对两个模拟研究结果的呈现顺序做了调整，统一先呈现项目参数估计结果，然后呈现潜变量估计结果。在潜变量估计结果中统一先呈现被试个体能力值估计结果，再呈现潜变量方差估计结果。只是由于两个模拟研究内容和目的的不同，结果呈现方式有所差异，本次修改在文中在呈现结果之前给出了明确的解释和说明。研究一中修改内容见 3.1.4 节中红色字体的内容，主要调整原文中图 5 和图 6 的呈现顺序，以及结果解释的位置和流畅性。研究二中修改内容见 3.2.2 节中红色字体内容，结合意见 3 和意见 6 对结果内容和呈现方式都进行了修改。

意见 6：在进行仿真研究时，可考虑与 Two-tier 模型进行对比，这样才能体现出 MTRM 模型的优势，特别是对研究二的结构 2。

回应：

感谢审稿专家的意见。本文 MTRM 和 two-tier 确实有相近的地方，也各有自身的特点和优势，这一点在意见 3 的回复中进行了详细叙述。结合您给出的意见 3，本次修改在模拟研究 2 中的结构 1 和结构 2 中，同时加入 two-tier 模型估计结果，并与原有的结果进行对比，讨论 MTRM 和 Two-tier 模型之间的区别以及各自的优势。主要修改内容见文章 3.2.2 节中红色字体的内容。

意见 7：文中绘制的图建议不要边框，图 4、5、6、8、10。

回应：

感谢审稿专家的意见。本次修改对文中所有图边框进行调整和修改，修改方式为取消了外边框，添加内部坐标轴线，修改后图片分别见图 4、5、6、8、9。

意见 8：表 2 中的"阈值"指什么？上下文中没有交代。

回应：

感谢审稿专家的意见。原文中的"阈值"原本出现在模拟研究 2 中，之前没有进行交代。本次修改对模拟研究 2 进行了较大修改，包括加入 two-tier 模型、调整结果呈现顺序和呈现方式等，修改后全文统一呈现多级计分项目的项目步骤难度估计结果，不再出现原来的"阈值"，修改后对文中新的结果呈现方式也做了更详细的描述和介绍。

意见 9：文中存在一些重复信息，如 1.1 节第 1 自然段中“被试在同一题组下（例如：同一篇阅读短文）各个题目的作答反应会受到相同阅读材料和背景的影响。”和上面的话“例如在英语水平测验中，一篇阅读材料往往会包含多个题目。”表达相同的意思。第 3.1.3 节中的话“通过偏差（bias）、绝对偏差（MAE）、均方根误差（RMSE）以及参数估计值和真值之间的相关（Corr）”在上一自然段中刚刚出现过。2.3 节中“中使用 MCMC 方法实现 conquest 对模型 MTRM 的参数估计”，让读者很难准确理解作者想表达的意思。

回应：

感谢审稿专家的仔细阅读。原文部分表述不够准确流畅，本次对文中多处描述都进行了修改和完善，使整篇文章在表达上更加流畅和清晰。修改内容包括但不仅限于本条意见中提到的部分。如：

对 1.1 节第 1 自然段的内容进行删减和调整，修改后的内容见 1.1 节第 1 自然段红色字体内容。

原文的 3.1.3 节现为文章中 3.1.4 节，原文中二次出现的话“通过偏差（bias）、绝对偏差（MAE）、均方根误差（RMSE）以及参数估计值和真值之间的相关（Corr）”已经删除，见 P9 中 3.1.4 节。

对于“使用 MCMC 方法实现 conquest 对模型 MTRM 的参数估计”，这句话想表达的意思是，使用 conquest 进行参数估计，因为 conquest 在估计过程中可以选择不同的参数估计方法，包括极大似然估计和 MCMC 等，本文选择的是 MCMC 参数估计方法。文中对这句的话表述稍微进行了调整，见文章 P7 中 3 章前一段红色字体。

另外为了使文章表达更流畅更清晰，减小重复信息的出现，本次修改对原文 2.1 节前面大段的过渡内容进行了精炼和完善，并调整位置放到 1.2 节问题提出的第一段，见 P3 中 2.1 节第一段红色字体。

第二轮

审稿人 1 意见：

意见 1：

文中批注：新题目依旧不合适，新模式是对原有模型的约束，如果没有限定使用情境，为什么能用拓广或拓展这样的词？不知“多维题组反应模型：多维随机系数多项 Logistic 模型的应用拓展”如何？仅供作者参考。另外，摘要和正文等地方的描述也需要修改

回应：

感谢审稿专家的意见，确实增加“应用”两字后使得文章题目更加得当，因此本文进行了修改，另外文中相应地方的描述也进行了修改，修改内容都用红色字体标出。

意见 2：

文中批注：原文中“目标能力的多维性包括项目间多维和项目内多维”需引用 Adams 等(1997)

回应：

感谢审稿专家专业和严谨的态度，已经进行了引用。修改内容见修改稿第 4 页红色字体。

意见 3：

文中批注：原文中“Adams 等（1997）将单维随机系数多选逻辑斯特模型（RCMLM）进行拓展，得到可以应用于包含多维能力测验的多维随机系数多项逻辑斯特模型（MRCMLM）”

这句话上面已经提到了，没有必要重复，直接说 MRCMLM 可描述为　即可

回应：

感谢审稿专家的意见，这里已经按照专家建议进行了简化。修改内容见文章第四页红色字体。

意见 4：

文中批注：请使用矩阵符号。

回应：感谢审稿专家严谨的学术态度，已经对原文中的矩阵符号进行了修改。

意见 5：

文中批注：原文中“在本研究中，为了简化模型，固定所有潜变量的载荷为 1，但在实际中，多维测验情况往往更加复杂，一个项目在不同维度上的载荷大小可能有所不同，因此对潜变量载荷的进一步探讨也是有必要的。” 这是 Rasch family models 的基本假设。当对模型引入区分度参数后，会导致逻辑上的矛盾。比如，某一份测验得到结论是 考生 A 的能力比考生 B 高，这个结论从逻辑上表明：考生 A 在作答这份测验的任何题目时的正确作答概率都应该高于考生 B，否则就不能称其能力高于 B。而这种逻辑正确性仅在 Rasch modeling 下是可维持的，引入 slope 后就会违背这种逻辑。另外，引入 slope 后，从 ICC 可看出 slope 越高的题目，其 ICC 越陡，即对低能力区间内和高能力区间内的被试的区分度越低，违背了 slope 本身的意义。所以，这点不是不足，只是和另外一部分从因素分析角度考虑问题的作者们的观念上的差异。

回应：

感谢审稿专家的指导，确实这句表述不够准确，本次修改对原文讨论部分进行了完善，这句内容在修改后已经不存在。

……

审稿人 2 意见：

意见 1：作者对论文做出了比较大幅度的修改，进一步完善了研究的设计和论文撰写，这点值得肯定。但是，作为一篇开发新方法的论文，我们使用仿真实验和真实数据的分析，都必须围绕证明新方法的优势而开展的。无论是实验的设计、结果和讨论、以及摘要的撰写，都应该仅仅围绕这一核心内容展开，不宜把点放得过散。

回应：

非常感谢审稿专家的意见。确如专家所言，文章的撰写和论述应该围绕核心问题展开，本文第一次修改对研究设计部分做了较大的调整，却忽略了文章内容的整体性和链接，描述相对较散，前言和总结部分也没有进行相应的承接和修改。针对这一不足，本次对文中多处描述进行了调整。主要包括：

（1）1.2 节问题提出部分，增加对 two-tier 模型的描述，更好的引出本文研究问题，并精炼和完善原来的表述，使得语言更加简洁明了。同时为避免内容重复和内容繁杂，删除原文 3.2.2 节中相应的对 two-tier 模型的细节描述，并精炼和完善原来的表述。相应修改内容见 1.2 节和 3.2.2 节中红色字体。

（2）第 4 节讨论和总结部分，删除原来过于细节却与核心问题关系不大的描述以及与前文重复雷同的内容，并修改原来对 MTRM 和 two-tier 模型的总结性话语，使语言更加准确简洁。相应修改内容见第 4 节红色字体。

除上述两个主要部分外，还对文中第 2、3 节中多处细节进行了调整，让文章的描述更加整齐集中，语言更加简洁准确。相应修改位置都用红色字体标出。

意见 2：目前“研究三：多维题组反应模型的应用研究”所使用的数据，不是项目内多维题组的情形。换个角度看的话，也就是作者自身都很难找到项目内多维题组的实际应用，会使得新方法的实践意义大打折扣。另外，作者通过研究三得出的主要结论是题组效应不应被忽略，这个结论是早已知晓的问题，与新方法的开发关系不紧密。研究三没有再使用 two-tier 模型，也是一个不足之处。

回应：

非常感谢审稿专家的意见，的确如评审专家所指出的，如果找不到项目内多维题组的实际应用，会使新方法的实践意义打折扣，因此本次对这一问题进行了修改。

原文之所以在前面的稿件中没有涉及项目内多维题组结构，并不是因为难以找到项目内多维题组的实际应用，而是因为使用数据的保密限制。根据专家的意见，在本次修改稿中的研究三里，我们冲破重重阻力，最大程度的获得了题目的原始信息。因此本次修改包含了项目内多维题组结构。经过对不同模型的分析和比较，我们发现在同时考虑项目内多维题组的影响下，模型拟合度提高，恰好验证了新方法具有重要的实际应用价值。修改后研究三的主要结论是任何可能存在的项目内多维题组效应都不应该被忽略，MTRM 具有应用价值。围绕这一核心结论，本次修改同时对研究三的呈现进行了简化，删除了不必要的结果呈现。该部分修改内容见 3.3 节红色字体以及 1.2 节问题提出部分的问题 4，和第 4 节总结中相应的红色字体。

本次修改研究三中并没有增加 two-tier 模型的估计结果，因为 two-tier 模型和 MTRM 参数估计结果并不在同一量尺上，两个模型也没有共同的模型拟合指标。总体来说，两个模型在实际应用中难以用相同的标准进行比较。因此一方面为了避免文章篇幅不必要的增长，另一方面为了避免上面提到的“散”的问题，不考虑在研究三中增加 two-tier 模型的分析结果。为了更好的让专家理解该数据在 two-tier 模型下的表现，特附上 two-tier 模型潜变量方差估计结果（如下图）。其中潜变量 1-3 对应数学的三个能力维度，4-10 依次对应文章中描述的题组 1-7。

Group Latent Variable Means:

Group	Label	mu 1	mu 2	mu 3	mu 4	mu 5	mu 6	mu 7	mu 8	mu 9	mu 10
1	G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Latent Variable Variance-Covariance Matrix for Group 1: G

Theta 1	Theta 2	Theta 3	Theta 4	Theta 5	Theta 6	Theta 7	Theta 8	Theta 9	Theta 10
2.73									
2.64	2.88								
1.75	1.80	1.80							
0.00	0.00	0.00	1.01						
0.00	0.00	0.00	0.00	1.02					
0.00	0.00	0.00	0.00	0.00	0.98				
0.00	0.00	0.00	0.00	0.00	0.00	0.99			
0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00		
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97

意见 3：文中的细节还需要再检查，例如有些不必要的空格，特别是表 2 和表 3 中；三线表的表格线粗细不符合常见形式；英文 unidimensional 被拼写为 undimensional 等。

回应：

非常感谢审稿专家的意见。本次修改对文章细节问题进行了检查和修改。三线表按照 APA 格式要求“顶线和底线为粗线，排版时俗称‘反线’；栏目线为细线，排版时俗称‘正线’”进行了粗细的调整。英文拼写也进行了修改。另外，专家指出的表 2 和表 3 中空格不知是否

指的是表格下方多余的空行，本次修改对文中不必要的空行进行了调整。

再次感谢专家费心审阅此文，专家严谨的学术态度让我们非常钦佩，专家的每一点意见对我们都是莫大的帮助。

第三轮

审稿人 2 意见：

意见 1：经过这两轮的修改，论文的质量得到提高，推荐发表。请在文中解释清楚表 5 中的 Final Deviance 代表什么，如有可能改为中文术语。

回应：

非常感谢审稿专家对本文的修改给出的建议和督促。文中已经进行了相应了修改，根据 ConQuest 手册中对 Final Deviance 的描述添加了说明。因为手册中对于该指标的描述更多的从可以作为模型拟合指标进行模型比较方面进行，为了保证修改内容的准确性，严格按照手册中的描述进行说明，翻译为模型方差，并进行引用。具体修改内容间第 14 页红色字体。