《心理学报》审稿意见与作者回应

题目：多维题组效应 Rasch 模型
作者：詹沛达 王文中 王立君 李晓敏

第一轮

审稿人 1 意见：本文对题组概念进行了重新定义，突破了以往对于题组概念的认识，并结合已有方法，提出了可处理项目内多维题组效应的多维题组效应 Rasch 模型，成功地解决了二值计分和多值计分问题，在理论和实际应用中有较大创新。几点修改建议如下：

意见 1：引言部分需直奔研究主题，作者运用大量篇幅介绍题目及传统处理方式的不足，太啰嗦，应该突出本研究的意义，读起来让人感觉太累，无法一下子抓住研究重点。建议：第二段可以考虑删除，因为第一段已经介绍什么是题组了，2.1 部分也有介绍。第三与第四段可压缩为一段，先介绍传统方法缺陷然后引出题组的处理方法。第五段才是本文的重点，要着重引出为什么要关注“多维”的重要性及意义，因为“多维”是本文的最大卖点，其他部分可以简略些。

回应：感谢您提出的宝贵意见，我们已经采纳您的建议，将原文第 2、3、4 段进行了删减、压缩与合并，在保证文献综述完整性的前提下共缩减约 720 字，相较于原文更凸显主题。

意见 2：“2.3”与“2.4”部分内容与参考文献 Wang，& Wilson（2005c）内容基本一致，有很多雷同地方，在方法上没有太大创新。

回应：我们认为 IRT 领域内的研究，模型的建立与改良是为解决实际应用问题而服务的。当测验中出现一个未被现有模型(如：Rasch 题组模型(Wang & Wilson, 2005c))包含在内的潜因素(如：项目内多维题组效应)时，其中一个有效的做法是将其纳入现有模型或构建新的模型。

本研究的目的是建构一个可以处理项目内多维题组效应的 IRT 模型。我们所要解决的问题(如何处理测验数据中包含的项目内多维题组效应？)与 Wang 和 Wilson(2005c)所要解决的问题(如何处理测验数据中包含的项目内单维题组效应？)是具有递进关系的。因此我们选择在原有模型的基础上进行扩展。由于扩展后的模型是基于原模型开发的，所以在描述两者的建构过程时难免地会出现雷同之处，但这并不影响拓广模型对新问题的处理能力。

意见 3：“3.1.3”部分，多级评分项目均设定为包含 3 个分数级别(0 分,1 分,2 分)，且各项目的第 1 临界参数分布满足 τ_{i1}~U(-1,0)。为何要规定各项目的第 1 临界参数分布满足 τ_{i1}~U(-1,0)这一条件，在给定 bi 和 bij 后，tij 的值是确定的。请作者稍作解释。

回应：您说的对，因为在 PCM 或 GPCM 中有 $b_{ij} = b_i + (b_{ij} - b_i) \equiv b_i + \tau_{ij}$，所以 b_i、τ_{ij} 和 b_{ij} 这 3 个参数只要给定了其中 2 个，那么另一个就是定值。原文中不是“规定”而是“描述”了我们生成项目参数时的情况，且这与模拟被试作答时生成参数的顺序有关，我们是先生成定位参数 b_i 然后再根据所需的分数级别来生成 τ_{ij} 的，而 b_{ij} 是由前两个参数来确定的。但不同的研究者生成这 3 个参数的顺序可能不同，本研究中我们直接给出了研究中 3 个参数的分布，且 3 个参数分布关系也是合理的(极值点相加减满足 $b_{ij} = b_i + \tau_{ij}$)

此外，ConQuest 软件也仅提供 b_i 和 τ_{ij} 参数的参数估计结果而不提供 b_{ij} 的，因此我们的

实验结果(表 2 和表 4)中也只给出了对 b_i 和 τ_{ij} 参数估计的返真性，而这就需要在参数设定时对 b_i 和 τ_{ij} 参数的分布进行交代。

意见 4：文中有很多句子前后重复，篇幅稍显过长，请作者精简语言。

回应：感谢您提出的宝贵意见，我们在后期阅读中也意识到原文篇幅稍长，这与我们初期想把项目内多维题组效应给读者讲清楚有关。我们已经采纳您的建议，对原文进行删减、压缩与合并，全文共减少约 4000 字，可读性大幅度增加。

意见 5：参考文献顺序请检查，有错误。

回应：谢您的细心审阅，由于正文部分出现改动，我们已经重新校对了参考文献。

意见 6：英文摘要中，一般不出现公式，请作者斟酌。

回应：感谢您的建议，我们已对英文摘要进行修改。

意见 7：如何理解项目内单维题组效应属于项目内多维这一说法？

回应：由于本研究的关注点在题组效应的维度上，所以并没有强调被试能力的维度问题。“项目内多维(度)(Within-item multidimensionality)”这一概念是指单一项目反应内包含多维测验潜因素，是项目水平的描述。而“项目内单维题组效应”仅指的是单一项目反应内包含了一维题组效应，而该项目反应内还受被试能力维度的影响，因此从项目水平看项目内单维题组效应也属于项目内多维度。

意见 8：原文 2.4 内容出现在“2 多维题组效应 Rasch 模型的开发”部分稍显奇怪。

回应：结合两位审稿人的意见我们对原文章节安排进行了调整，新增加了“3 项目内多维题组效应 Rasch 模型的参数估计”一章，并整合了原文中“2.4 MTERM 与 MRCMLM 间的转换”中的内容，且添加了一些内容，具体请参见修改稿。

意见 9：据本文定义，项目内多维指的是题组的维度数，是否应该将能力维度包含在内？

回应：项目内多维指的是单一项目反应内包含多维测验潜因素。“项目内多维题组效应”指的是题组效应的多维度性，而“项目内总维度”指的是题组效应维度+被试能力维度，因此我们仍需将能力维度也包含在内。

审稿人2意见：本文的选题具有实际和理论意义。文章基本达到了心理学论文的规范。本文的可取之处在于对研究的细节介绍得很清楚、详尽，而且通过举例子的方式试图促进读者的理解。然而，本研究有一些比较大的问题，希望引起作者的注意。

意见 1：有没有必要扩展“题组”这个概念的含义？正如作者所说，“相同的单元内容(如：数学测验中考查“四则运算”单元的所有项目)、相同的题型(如：英语测验中所有的听力题)、相同的知识点(如：科学测验中所有考查“重力”的所有项目)、相同的先验猜测概率(如：数学测验中所有的四则一选择题)甚至是评分者(如：测验中某评分者所审改的所有项目)”都受到局部独立性的影响。我相信所有熟悉 IRT 的研究者都熟知这一点。我们已经有了局部独立性这个概念，为什么还要违背人的常识，去拓展“题组”这个概念？“相同题型”就是“相同题型”，人们从字面上就能理解，如果非要把这也叫做“题组”，让人不太容易接受。这样对我们认识这些现象有什么好处？如果没有特别的必要，不如不要提这个新的概念——不但没有益处，反而增加认知负担、造成误解。

回应：感谢您提出的问题，我们相信国内有部分研究者也会存在和您一样的疑问，为使读者

更易于理解，我们对原文中的表述进行了适当的修改。这里针对您的问题我们给出更详细的解释：

首先，如正文所述，我们从相关文献中提取出题组概念的两个核心元素分别是 共同刺激和项目集合，并将两者结合进一步指出题组的本质是“一个存在共同刺激的项目集合”。与之前研究相比，我们扩展的并不是题组的概念，而是先明确了题组概念的内涵进而扩展(填补)了题组概念的外延和应用范围。原有的应用范围过于狭窄或单一，概念的外延和内涵并不互补。不尝试去突破已有“常识”和思维定势，何来创新一说呢？

其次，之所以把某项目集合(如“相同题型”)称为题组，是因为被试对该项目集合的反应受到测验目标潜质外的某些共同刺激的影响，违背了狭义局部项目独立性假设。以“相同题型”为例，我们之所以将“相同题型”归为题组，也是因为在实际测验中不同考生对不同题型存在不一致的认知和作答倾向(也可将其看成一种(干扰)潜质，但并不是该测验的目标潜质)，比如在 IELTS 的 Listening 和 Reading 中，“老手”考生倾向于作答填空题，而“新手”考生却倾向于作答选择题，再如不同考生群体对不同分值的项目也存在倾向性差异(Ariel, Dunlosky, & Bailey, 2009; Dunlosky & Ariel, 2011)，或者不同题型的先验猜测概率存在差异(Leong, Ling, & Mahdi, 2012)等等。而往往这类干扰潜质会影响后期的参数估计，导致高估测验信度、项目参数偏差估计、增加测验等值误差等诸多问题。因此本研究旨在处理题组效应，通过扩展模型将其从目标潜质中抽离出来(如何处理项目内多维题组效应)。

最后，我们之所以使用“题组效应”而不是“局部项目依赖性”，是因为题组效应只是局部项目依赖性中的一种，还有被试群聚效应等其他潜因素可产生局部项目依赖性。

意见 2：公式 10 实际上是标量和向量之和，是无法相加的。希望作者重新表示自己的公式。

回应：非常感谢您发现该错误，这是我们在论文写作时的疏漏，对此我们已经将原公式的表述方式进行了修改，使其和后文“3 多维题组效应 Rasch 模型的参数估计”一章内的推导公式(或原稿中公式(13)、(14))保持一致，详情请参见修改稿。

意见 3：本研究似乎提出了新的模型，但实际上是用 conquest 软件，对多维随机系数多项逻辑斯蒂克模型的一个应用。那么，这个模型新在何处？此外，本研究提出的实际问题，已经有模型可以解决，请参考 Cai, L. (2010). A Two-Tier Full-Information Item Factor Analysis Model with Applications. *Psychometrika*, *75*(4), 581–612.。因此，希望作者重新审视本研究。

回应：感谢您的提问，我们已在修改稿中增加“3 多维题组效应 Rasch 模型的参数估计”一章。修改后的文章整体结构更加明确，对表述不清晰的地方也进行了修改。这里针对您的问题我们给出更详细的解释：

首先，本研究的目的是建构一个可以处理项目内多维题组效应的 IRT 模型。模型是为解决问题而服务的，直接对已有模型进行多维度拓广是一个较为有效的途径。从模型拓广思想看，新模型可以看成 Rasch 题组模型的拓广模型，参数估计可以经由很多现有的统计软件实现，例如 ConQuest, WinBUGS, R 以及 Matlab 等等，而在这其中，采用 ConQuest 软件来实现参数估计是一个较为直接且有效的选择。

其次，多维随机系数多项逻辑斯蒂克模型是一个宏观的、开放式模型，是 ConQuest 软件实现参数估计的基础，如果直接使用该模型来处理现实情景就显得过于夸张和繁琐。比如在处理二级评分数据时，实际测验分析人员会倾向选择简单易行的二级评分模型（e.g., Rasch）,而不是多级评分的模型(e.g., PCM)，更不会采用多维随机系数多项逻

辑斯蒂克模型,而前两者恰恰都是多维随机系数多项逻辑斯蒂克模型的约束模型(或是您说的“一个应用”)。与此类似，IRTPRO 软件是以2-tier 模型为基础建构的，2-tier 模型也是一个宏观模型，直接使用2-tier 模型来处理现实情景必然也要给出繁杂的约束条件(尤其是软件说明书中并未涉及的测验情景)。但由于版权问题我们并没有使用过该软件，因此暂不清楚 IRTPRO 是否可以或如何添加约束条件来实现处理项目内多维题组效应(目前已知 IRTPRO 可处理单维题组效应)。其实，以上这些模型都是广义线性模型(generalized linear model)的特例,ConQuest 和 IRTPRO 所能解决的问题，通过对广义线性模型进行约束设定大都可以解决，但我们不能说这些模型的提出是没有意义的，因为它们简化了广义线性模型，为实际使用带来了便利。

最后，本研究的创新点不在于模型或参数估计的开放，而是指出了项目内多维题组效应这一在以往测验分析中被忽略的问题，並建构解决该问题服务的模型。而至于采用什么软件和方法来实现参数估计，并不限于 ConQuest 软件，也可由测验分析人员自己选择。

意见 4：本文通过举例子的方式来介绍作者的思想，是可取之处。但一些例子，如“女人和她的丈夫因婚姻关系可以组成一个家庭，同时，这个女人还可以和她的父亲因血缘关系组成另一个家庭。或许，她还可能认养了另一个无血缘关系的小孩。”似乎不太适合的篇幅有限的学术论文中出现。而更适合在讲座和书中出现。

回应：我们已经采纳您和第一位审稿人的建议，以提升全文质量和可读性为原则，对原文中的赘述的、不恰当的、过于繁杂的描述进行了大幅度删减、压缩与合并，全文共减少约 4000 字，可读性大大增加。

意见5：本文中有大量的模型名称及缩写，例如：

标准项目反应模型(Standard Item Response Model, SIRM)

基于题组的计算机化自适应测验(testlet-Based Computerized Adaptive Testing, TBCAT)

标准题组模型(Standard TRM, STRM)

单维题组效应模型(Unidimensional testlet-Effect Model, UTEM)

单维题组效应 Rasch 模型(Unidimensional testlet-Effect Rasch Model, UTERM)

单维题组效应双参数模型(Unidimensional testlet-Effect Two-Parameter Model, UTE2PM)

单维题组效应高阶模型(Unidimensional testlet-Effect Higher-order Model, UTEHM)

多维题组效应 Rasch 模型(Multidimensional testlet-Effects Rasch Model, MTERM)

二级评分的 MTERM(记为 dMTERM)

多级评分的 MTERM(记为 pMTERM)

多维随机系数多项逻辑斯蒂克模型(Multidimensional Random Coefficients Multinomial Logit Model, MRCMLM)

缩写本来是为了减轻人的认知负担。例如 IRT 这个缩写，大家都知道是什么，说 IRT 就比项目反应理论更有效率。但这些不那么广为人知的模型（而且名字非常长），又大量地以缩写的形式表达出来，使得文章的可读性大大下降。希望作者谨慎提出模型的名字以及谨慎使用缩写。

回应：非常感谢您的建议，我们在后期阅读中也发现原稿英文缩写过多的问题，对此我们已大幅度删减非本文主要涉及的英文缩写，如：STRM、TBCAT、UTE2PM、UTEHM 等，保留了 MTERM(本研究提出的模型)、TRM、PCM、RSM、GPCM 等以及表格中的缩写。修改后，全文的可读性的确大幅度增加！

第二轮

审稿人 1 意见：修改稿质量明显高于第一稿，基本符合心理学报发表要求。

审稿人 2 意见：作者较好地回答了审稿人的问题，根据审稿人的意见进行了有针对性的修改。但是，在上次评审意见中建议作者参考 two-tier 模型，并非建议作者参考 IRTPRO 软件，希望作者对模型和软件有所区分。软件只是实现模型参数估计的工具。建议作者再对一些语言进行修饰和润色。例如，在总结和展望部分的第一段末，作者直接指出“本文的创新点是……”。根据我本人的理解，研究的创新点可以留给读者自己来发现，一般无需在文章中直白地“自我宣传”，但这也许是行文的风格，仅供作者参考。