《心理学报》审稿意见与作者回应

题目：基于分部评分模型思路的多级评分认知诊断模型开发
作者：高旭亮; 汪大勋; 王芳; 蔡艳; 涂冬波

第一轮

审稿人 1 意见：
意见 1：稿件描述了分部评分模型下认知诊断的模型构建研究, 对于拓展 CDM 的应用领域还是很有价值和意义的,但是文中新模型应该有名字, 不建议笼统用新模型来代替。
回应：非常感谢专家的意见, 根据专家的意见, 我们已经使用 GPCDM(新模型的英文简称)代替了新模型的表述。

意见 2：稿件中述及等级评分模型有累积概率、连续比率和相邻类别三种模型, 并且前两种的认知诊断模型已有不少, 基于相邻类别模型的CDM有何独特的价值和意义, 请予以阐明。
回应：感谢专家的宝贵建议, Mellenbergh (1995)根据模型将多级评分数据二级化的方式将 IRT 的多级评分模型分为 3 类, 这 3 类模型将多级评分数据二级化的方式是完全不同的。假设题目满分是 3 分, 定义 , 累积概率模型(cumulative probability models)二分为 和 , 而连续比率模型(continuation-ratio models)则二分为 和 , 相邻类别模型(adjacent category models)二分为 和 。因此, 这 3 类模型的建模思路是完全不同的, 各有特点, 累积概率模型侧重于分析某个等级以上(包括该等级)所有等级与该等级下(不包括该等级)所有等级之间的关系; 连续比率模型侧重于分析某个等级以上(包括该等级)与该等级的以下一个等级之间的关系; 而相邻类别模型侧重于分析两个相邻类别之间的关系, 该类模型可以深入分析每个解题步骤的加工信息, 即可以深入学生的解题过程, 因而它对于深入探讨被试的加工过程更有优势, 也有望提供更为丰富的诊断信息。另外, 我们在正文中重新对这 3 种多级评分模型的各自特点进行了阐述, 并补充了基于相邻类别模型的 CDM 优势。

意见 3：稿件中 P7“βjx, uv , 表示 αlu 和 αlv 的交互效应”应写成“……的二阶交互效应”, 后面一句最好写成“……K 阶交互效应”, 此外, 公式 6 上面一行的交互效应应是最高阶的交互效应, 如果是交互效应, 应该包含从二阶交互到最高阶交互的所有项。
回应：谢谢专家的意见, 我们已经根据专家的意见进行了修正。

审稿人 2 意见：不同于以往分部评分类多级评分 CDM(比如 GDM 和 PC-DINA)将 Q 矩阵定义在题目水平以及“要么只考虑属性的主效应、要么只考虑属性间的交互效应”, 本研究基于分部评分模型思路提出的新模型(即 GPCDM)不仅将 Q 矩阵定义在得分类别水平(即 Cat-Q; 在此基础上容易合成题目水平的 Q 矩阵), 而且能够同时考虑属性的主效应和交互效应。因此, 本研究相对于已有文献具有一定的贡献。另外, 本文的文献综述比较全面, 逻辑层次结构比较清晰。文章存在的不足总结如下:
意见 1：第 3 页“中文摘要”部分: (1)没有必要重复强调“新模型具有重要的理论意义和实践价值”; (2)“新模型的参数估计精度指标达到了理想水平”, 什么是理想水平？(3)建议按照学报的要求撰写这部分。

回应：非常感谢专家的意见，根据专家的意见，我们对摘要部分进行了修改。
重新修改后的摘要如下: “基于分部评分模型的思路，本文提出了一般化的分部评分认知诊断模型(General Partial Credit Diagnostic Model, GPCDM)，与国际上已有的基于分部评分模型思路的多级评分模型 GDM(von Davier, 2008)和 PC-DINA(de la Torre, 2010)相比，GPCDM 的 Q 矩阵定义更加灵活，项目参数的约束条件更少。Monte Carlo 实验研究表明，GPCDM 模型的参数估计精度指标 RMSE 介于[0.015, 0.043]，表明估计精度尚可； TIMSS(2007)实证数据应用研究表明，与 GDM 和 PC-DINA 模型相比，GPCDM 与该数据的拟合度更好，并且使用 GPCDM 分析该数据的诊断效果也更优。总之，本研究提供了一种约束条件更少、功能更为强大的多级评分认知诊断模型。”

意见 2：第 4 页倒数第 2 段中的“logit”翻译为“对数”可能不太合适。
回应：感谢专家指出，我们查看了已发表的文献对于“logit”的表述方式，发现几乎都是直接使用“logit”的本身形式，并没有翻译成中文表达式，例如，以下文献中都直接使用了“logit”来表达:
[1] 丁树良，吴锐，张节兰，熊建华. (2008).概率分布等值法及其应用.心理学报, 1, 101–108.
[2] 李付鹏，宋吉祥，杜海燕，赵发忠，储林林. (2019). 基于 Rasch 模型的高考数学性别 DIF 检验.中国考试, 3, 43–47.
[3] 詹沛达，边玉芳，王立君.(2016).重参数化的多分属性诊断分类模型及其判准率影响因素. 心理学报，48(3)，318–330. 因此，为了表述的准确，我们在正文中也改用了“logit”的形式来表述。

意见 3：第 4 页倒数第 2 行: 什么是“基于等级反应的理论提出了”?
回应：感谢专家细致审稿，是我们的表达不准确，我们已经将这句话修改为“基于等级反应模型(GRM)的建模思路”。

意见 4：第 6 页“2 基于分部评分模型思路的多级评分 CDM 开发”部分第 1 段第 3 行: “属性模式掌握某个属性”的说法不准确。

回应：感谢专家指正，我们将这句话改为“属性模式为$\boldsymbol{\alpha}_l$的被试掌握了第k个属性”。

意见 5：第 9 页公式(11)下面第 1 行: 在使用 optim 函数计算 M 步的极大值时，有没有什么特别的参数设置需要强调或说明?
回应：我们在文中对 optim 函数的使用进行了说明，具体如下: optim 函数在 R 里表达式为 optim (par, fn, method), par 代表项目参数初值, fn 代表目标函数, method 选择优化算法。在使用 optim 函数计算极值时需要输入 par(项目参数初值)，本研究的初值从均匀分布中随机生成, fn(目标函数，即公式 11)，和选择的优化算法(拟牛顿算法)即可。

意见 6：第 9 页“4 实验 1: Monte Carlo 实验研究”部分第 1 段第 1 行: 实验 1 如何能够检验 GPCDM 模型的合理性与科学性?
回应：感谢专家的建议，这是我们行文表达有误，根据专家意见，我们调整了这句话的表达: “实验 1 的目的是检验 GPCDM 模型的参数估计精度及其性能”。

意见 7：第 9 页倒数第 2 行: 这样设计 Cat-Q 的理由是什么？请给出解释。
回应：谢谢专家的建议，我们对 Cat-Q 的模拟主要是借鉴了 Ma 和 de la Torre(2016)研究中的

做法: 多级评分的题目中每个得分类别最多考察 2 个属性, 并且 Cat-Q 中每个属性的测量次数都是相同的。另外, 为了提高诊断测验的效果, 5 属性和 7 属性的 Cat-Q 分别包含了 5 个和 7 个二级评分的题目且这些项目包括一个完整的可达矩阵(R 阵)。

意见 8：第 9 页倒数第 1 行: 40 题和 50 题时的 Cat-Q 与 20 题和 25 题时的 Cat-Q(如表 2 和表 3 所示)是什么关系？是重复的关系还是其他关系, 需要交代清楚。
回应：谢谢专家的建议, 40 题和 50 题时的 Cat-Q 与 20 题和 25 题时的 Cat-Q 是重复关系。另外, 我们在正文中增加了 40 题和 50 题时的 Cat-Q 与 20 题和 25 题的关系说明。

意见 9：第 11 页“4.1.1 被试参数的模拟”部分: 缺少对模拟被试数的描述, 即 N = ？。
回应：感谢专家指出, 我们在实验最开始自变量的描述中已经包含了样本量的设置方式。但为了让读者更加明确, 根据专家的建议, 我们在“4.1.1 被试参数的模拟”部分增加了样本容量的说明, 即 N=500, 1000, 2000, 4000 四个水平。

意见 10：第 12 页第 1 段: 作答如何从对应的分类分布中抽取？需要描述准确、描述清楚。
回应：感谢专家的建议, 为了读者更好的理解, 我们进行了举例说明: 假设被试在某一题恰得 t 分(满分是 4 分)对应的概率分别是{0.03, 0.08, 0.12, 0.63}, 则被试在该题的得分从{0, 1, 2, 3, 4}中抽取一个数, 而每个得分被抽取的概率分别是 0.03, 0.08, 0.12, 0.14 和 0.63。这种模拟方法也是 Ma 和 de la Torre (2016)研究中所采用的方法。感谢专家的指出。

意见 11：第 12 页公式(12)和(13)下面一行公式项都表示“第 r 次实验”而非“第 R 次实验”。相比公式(13)所呈现的综合的 RMSE 结果, 我更想看到各个题目参数层面的 RMSE 结果。
回应：感谢专家的建议, 按照专家的建议, 我们补充了每个题目参数的 RMSE 结果, 由于每个实验条件下题目层面的 RMSE 有相似的趋势, 因此, 限于文章篇幅的原因, 我们只截取报告了“属性个数等于 5, 样本容量为 1000, 测验长度为 20 题”实验条件下每题的 RMSE 指标。增加的实验内容如下:

表 6 显示了在属性个数为 5, 样本容量为 1000, 测验长度为 20 题时, Cat-Q 和 Item-Q 条件下每一题的 RMSE 指标, 由于其他实验条件下的结果和表 6 有相似的趋势, 因此, 限于篇幅的原因, 本文只提供了一种条件下的结果。

表 6 当 K=5 和 N=1000 时每题的 RMSE 值

题目	**Q** 矩阵的类型		题目	**Q** 矩阵的类型	
	Cat-**Q**	Item-**Q**		Cat-**Q**	Item-**Q**
1	0.025	0.095	11	0.025	0.082
2	0.032	0.092	12	0.026	0.088
3	0.033	0.069	13	0.027	0.091
4	0.036	0.081	14	0.029	0.086
5	0.024	0.086	15	0.028	0.088
6	0.034	0.082	16	0.018	0.019
7	0.033	0.083	17	0.021	0.020
8	0.023	0.079	18	0.019	0.019
9	0.034	0.069	19	0.020	0.019
10	0.024	0.084	20	0.020	0.021

从表 6 的结果可以发现，由于后 5 题是二级评分的题目，此时 Cat-**Q** 和 Item-**Q** 是等价的，因此 Cat-**Q** 和 Item-**Q** 的 RMSE 值基本相当； 而在多级评分的前 15 题中，基于 Cat-Q 得到的 RMSE 值始终要小于基于 Item-**Q** 的 RMSE 值，基于 Cat-**Q** 的最大 RMSE 是 0.036。另外，还可以发现，二级评分题目的 RMSE 要略低于多级评分的题目，这是因为，二级评分题目考察的属性个数要少于多级评分的题目。这个结果充分表明，EM 算法可以提供精确的参数估计精度，和 Item-**Q** 相比，使用 Cat-**Q** 有助于提供更多有价值的诊断信息，从而提高诊断测验的精度。

意见 12：第 14 页第 2 段中的“表 3”和“表 4”，应改为“表 4”和“表 5”。
回应：感谢专家细致的审稿，按照专家意见，我们已经将 14 页第 2 段中的“表 3”和“表 4”，应改为“表 4”和“表 5”。

意见 13：第 15 页倒数第 1 行至第 16 页第 1 行：每个属性的什么的相关系数？请描述费舍尔 z 变换是如何进行的？
回应：非常抱歉，这句话是我们行文的疏忽，我们直接计算了 8 个属性的平均边际概率(表 8 最后一列的结果)，并没有涉及相关系数和费舍尔 z 变换的计算，即这句话是多余的，这是我们的疏漏，给审稿专家带来的困惑深表歉意。在正文中，我们已经将这句话删除。

意见 14：第 18 页第 2 段第 1 行，“DINA 模型”应改为“PC-DINA 模型”。
回应：谢谢专家的指出，我们已经在正文中进行了更正。

意见 15：文中只要是表示向量和矩阵的符号都应该加粗呈现，比如 Q 矩阵。
回应：感谢专家指出，我们仔细检查了文中的公式，对于向量和矩阵的符号用了加粗呈现。

意见 16：文中存在多处表述不通顺的地方，比如: (1)第 18 页倒数第 1 段； (2)第 19 页的第 2 段和第 3 段。

回应：感谢专家的建议，根据专家的建议我们调整了上述 3 段表达不通顺的地方，另外，我们也再次通读了全文，修改了表达不流畅的语句。

第二轮

审稿人 1 意见：

意见 1：摘要部分是写 TIMSS(2007)而在正文中又写 TIMSS 2007，建议统一。

回应：感谢专家指出，我们已根据专家建议统一修改为 TIMSS(2007)。

意见 2：作者在修回稿件中详细阐述了三种将多级评分模型简化为二级评分模型的思路及其具体区别，非常感谢作者的反馈。我关注的是这三种区别仅仅是数学上的技术细节上的区别还是在应用领域上不同？比如三种不同思路的CDM在具体应用上各自有没有哪些什么更擅长处理的问题领域？

回应： 感谢专家的问题。我们在正文中增加了一段话来介绍这 3 类模型的特点及其应用特征。

累积概率模型是从整体出发考虑模型的建构，这类模型更适用于分析不强调具体解题步骤的诊断测验，例如，写作水平测验。而连续比率模型和相邻类别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(consecutive steps)，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； 而相邻类别模型是基于一个局部步骤(local step)来建模，即被试在当前步骤的作答只和前一步有关，这类模型更适合分析相邻步骤之间具有依赖关系的题目。Tutz(1997)认为相邻类别模型更适合分析评定量表(rating scales)类型的题目，连续比率模型更适合分析解答过程包含一系列连续步骤的题目。

意见 3：该模型从本质上看仍属于单一策略模型，给每个得分变量定义了一个属性掌握模式。但多数问题往往是多策略解题模式，最好在研究局限性中有所体现。

回应： 感谢专家的宝贵建议。按照专家的建议，我们在文章“讨论和展望”部分增加了一段，如下:

本研究开发的模型假设考生的解题策略只有一种，但在实际应用中，同一道题目经常存在不同的解题策略。如果在诊断评估测验中考虑了被试解题策略的差异，这也有助于提供更多有价值的信息，从而提高诊断的精度(涂冬波，蔡艳，戴海琦，丁树良, 2012)。因此，开发多策略的多级评分 CDM 值得未来进一步研究。

意见 4：从模型看，该模型的参数设定属于全模型，对于一个有 K 个计分类别的题目，大约有 k-1 个对应的属性掌握模式，每个模式下除了要估计全部的主效应外，还要估计所有可能的交互效应项，参数都能估计出来吗？在模型设定上有没有什么特别的要求？

回应：感谢专家的问题。现以一个例子来说明包含的参数个数，假设，题目满分是 3 分，若得分类别为 2 分时(即 step 2)考察的属性向量是(1,0,0,1,0)，则需要估计的参数包括: 1 个截距(intercept)参数，2 个主效应(main effect)参数，1 个交互效应(interaction effect)参数，该步总共需要估计 个参数，即对于任意一个得分类别，总共包含 个参数， 表示当前得分类别考察的属性个数。事实上，在同样的 Q 向量下，我们开发的新模型 GPCDM 和序列加工 G-DINA 模型(seq-GDINA, Ma & de la Torre, 2016)包含的参数个数是一致的，两个模型都采用了 EM 算法来实现其参数估计，seq-GDINA 模型的参数估计目前可以通过 R 软件中的 CDM 包(George et al., 2016)或者 GDINA 包(Ma & de la Torre, 2016)来实现。模拟研究发现 GPCDM

模型的参数估计精度指标 RMSE 最大值都小于 0.05，这和 seq-GDINA(Ma & de la Torre, 2016) 的参数估计精度是相似的。表明 EM 算法估计程序的结果稳定, GPCDM 模型的项目参数返真性较好，并且在进行参数估计时对模型设定并没有特别的要求。

审稿人 2 意见：作者认真回复了审稿人提出的问题并对稿件进行相应的修改。文章在正式接收前，还有几个小地方需要修改:

意见 1：第 5 页第 2 行至第 3 行, “如果 q_jx 考察了第 k 个属性”, q_jx 表示向量，它怎么会考察某个属性？

回应：感谢专家指出，我们将上述这句话改为: “如果 q_jx 包含了第 k 个属性”。

意见 2：公式(9)下面第一句话，对后验概率 P(alpha_l|X_i)的描述不准确，它表示“被试 i 在属性模式 alpha_l 的后验概率”。

回应：感谢专家的建议，我们将这句话改为: “P(alpha_l|X_i)表示被试 *i* 在已知作答向量 X_i 时属性模式属于 alpha_l 的后验概率”。

意见 3：公式(10)下面第一句话，对 M-step 目的的描述也不够准确。M-step 为什么要计算以下目标函数？

回应：感谢专家指出，我们将这句话改为: “M-step 的目的是使目标函数极大化的条件下来估计项目参数”。

意见 4：第 8 页倒数第 4 行, “这些项目包括一个完整的可达矩阵”，项目本身怎么会包括可达矩阵？

回应：感谢专家的细致审阅，我们已将这句话中的“项目”替换为“测验”。

第三轮

审稿人 2 意见：经过作者的反复修改，稿件质量有较大提高。在 4.1.2 标题下第 5 行“……掌握得分 x 所必需的任意一种属性”改为“掌握得分 x 所必需的所有属性”，是不是更好些？供作者参考。

回应：已经按照专家的建议进行了修改。