《心理学报》审稿意见与作者回应

题目：使用验证性补偿多维 IRT (CC-MIRT) 模型进行认知诊断评估
作者：詹沛达　陈平　边玉芳

第一轮

审稿人 1 意见：

意见 1：本文展示了验证性多维 IRT 模型的双参数模型（CC-M2PLM）的认知诊断功能，主要包括 CC-M2PLM 的推导，并利用模拟研究证明 CC-M2PLM 能够很好得拟合一个重要的认知诊断模型线性 Logistics 模型（LLM）生成的数据；同时论文用 CC-M2PLM，LLM 以及 DINA 模型来拟合两个实际数据，数据拟合指标表明 CC-M2PLM 是拟合最好的模型。这些结果表明，CC-M2PLM 虽然不是认知诊断模型，但是它能够很好的执行认知诊断的功能。

本文具有两大意义。第一，这篇文章的理论新意是揭示了虽然两种理论潜在的假设不同，但是多维 IRT 与认知诊断模型具有高度的等价性。第二，从多维 IRT 的角度来处理认知诊断的实际优势是，多维 IRT 已经发展比较成熟，为测量领域所熟悉，有很多成熟的商业与非商业软件可供选择，因此便于实际操作人员的使用。

论文作者仍需要对以下问题做出解释或者进一步阐述：

第一，168-174 行中，作者认为“得分矩阵来描述题目与维度之间的关系…但无论是 CDMs 中的 C 矩阵、Q 矩阵，还是 MIRTMs 中的得分矩阵等，它们的功能均是用于界定题目与属性之间对应关系，因此均可被称为验证性矩阵(confirmatory matrix)，”。作者的这个论断有可能是正确的，但是必须指出的是 Q 矩阵是在潜变量上定义的，而得分矩阵是基于观察分数定义的（参加 Adams 1997 年论文第 3 页第二段对此的论述）。因此，作者用来得到验证性多维二参数模型的 Q 矩阵可能与得分矩阵在概念上还是存在差异的。

回应：您的这条意见非常好，让作者们对原文中的描述产生了一丝怀疑。为确定原文中描述的准确性，作者们重新阅读 Adam 等（1997）和 ConQuest 手册，并梳理了 Scoring Matrix 这个概念，简曰如下：

对于单维的 RCMLM 而言，由于不涉及区分维度的问题，矩阵 $\mathbf{b}=(b_{11},...,b_{ij},...,b_{IJ})'$ 仅描述了题目 i 与得分 j 之间的关系。比如，对某 3 级评分题目{0,1,2,3}而言，则 bi=(bi1,bi2,bi3)=(1,2,3)，这对应了 PCM 中每个作答 level 的概率(i.e., Pnij)中分子部分包含的

θ 的数量，比如 j=1，则分子中含 1θ ；j=2，则分子中含 2θ 。需要强调的是，原文对这个矩阵 b 称为 scoring function，而 scoring matrix 的概念出现在 MRCMLM 一文中。

当 RCMLM 拓广到多维的 MRCMLM 时，现有的仅描述题目与得分之间关系的 scoring function 就扩展为涉及到维度的 scoring matrix，该矩阵中元素 bijk 就涉及到了题目 i 是否考查了维度 k，当 J=1 时(i.e., 二级评分题时)，得分矩阵中元素就变为 bik，即一个仅仅描述题目与维度之间关系的验证性矩阵。即得分矩阵本身比验证性矩阵(e.g., Q 矩阵)功能更丰富

综上所述，原文中的描述是恰当和合适的，但为了描述更为准确，我们在修改稿中添加了一个限制词：“……使用得分矩阵(scoring matrix)来描述(二级评分)题目与维度之间的关系……”。

意见 2：对比公式 1（LLM）与 3（CC-M2PLM），两者高度相似，我是否可以理解为除了被试潜变量的假设不同，两者是完全一致的；而这个不同是文中论述的两者差异的最大不同，其他的差异是由这个差异衍生而来（例如项目参数是否与被试参数同在尺度上）。因此，我是否可以解读为，从公式的角度看，本文提出的 CC-M2PLM 与 LLM 除了被试变量的差异之外，没有任何差异。

回应：您的理解是正确的，如果仅从数学函数角度讲两者是完全一致的。而对于这个“完全一致”需要从两个角度看：

(1)正是 LLM 与 M2PLM 看起来非常像（仅差了 Q 矩阵的概念），才让作者想到了 MIRT 模型是否也具有认知诊断功能这一问题，即本文的出发点和主题。为使 LLM 与 M2PLM 具有可比性，我们对 M2PLM 添加了验证性矩阵，将其限制为了 CC-M2PLM，此时两者在公式上的完全一致仅仅是为后续两者进行对比提供了保证。

(2)另一个角度看，两者看起来一致并不是说两者是一个东西，因为其内在东西(潜在特质)不同。这也是两者可进行对比的前提条件，即如果完全一致也就不需要进行对比研究了。

意见 3：237-238 行，能够具体说明一下在这里是如何具体生成属性相关高中低的，是否使用了 tetrachoric 相关系数？ 还有， 被试参数生成中采用了多元正态的形式，是否需要采用类似像一般的认知诊断研究那样直接生成，而不是这种通过多元正态形式间接生成的方式。

回应：感谢您的建议，我们认为您这两个问题其实是一个问题，因为本文 true model 是 LLM，所以被试参数就是属性。因此，原文 3.1.2 中采用多元正态分布来生成存在相关性的认知属性。一般的认知诊断研究生成属性的方法无法保证属性之间的相关性。不过作者们疏忽了用

四分相关来统计相关系数，采用 R 软件中的 polycor 包对属性间相关重新计算，并在修改稿中标出。计算后的四分相关系数均高于原皮尔逊积差相关系数(0.1 左右)，但仍在低(＜0.5)、中(0.5~0.8)、高(＞0.8)范围内，因此无需重复分析数据。另外，实证研究一中对参数估计值的相关也重新计算了四分相关。

意见 4：95-101 行中讨论了传统观念中关于 MIRT 与 CDM 模型的适用情况的讨论，我希望作者结合 CDM 中关于粒度的讨论再论述的详细一些，以利于其他研究者对这个问题的阅读。

回应：感谢审稿人的建议。属性“粒度”这个问题其实是比较复杂的，在当前作者的知识范畴下，暂无法想到如何将“粒度”这个概念进行合理量化，现有的量化方法会使它与属性“数量”相混淆。因此，我们无法确定起于“粒度”的研究，最终的结果到底是由“数量”导致的还是“粒度”本身导致的。

另外，讨论属性粒度这个概念的前提应该实在离散变量下(所以原文用了“含义”而非“粒度”一词)：从逻辑上讲粒度大的属性应该包含粒度小的属性，所以当被试掌握了粒度大的属性(e.g., 分数加法)则应该也掌握该粒度包含的所有粒度小的属性(e.g., 公分母)，那么当假设大粒度属性包含 4 个小粒度属性时，则应有 $\alpha_{(大)}=\prod_{k=1}^{K}\alpha_{(小)k}$，即所有小粒度属性都掌握才能掌握大粒度属性。而该公式并不能直接应用于连续变量上，换个角度讲，在连续变量下，如果被试分数加法这一（大）维度值很高(e.g.,2.9)，那么其公分母、进位等（小）维度的值也应该很高，但很难像在离散变量情况下去量化（大）维度与（小）维度之间的关系。

最后，鉴于作者对该内容的思考和理解并不成熟，且添加至引言部分稍显唐突，仅在这里与审稿人做上述交流下，见谅。

审稿人 2 意见：

意见 1：作者在论文中探讨了多维补偿 IRT 模型的认知诊断功能。对已有的探索性多维 IRT 模型引入了验证性矩阵进而得到了可与认知诊断模型(e.g., LLM)进行比较的验证性多维 IRT 模型(C-MIRTMs)，并展示了两类模型的诊断结果之间的对应关系，本文的问题在于引入切点的概念，将连续变量转变为二分变量，而切点直接选取为零。这一做法缺少理论依据，首先，将连续变量转换为二分变量的做法是否合适令人怀疑，其次，作者直接将切点取为零的理由是什么？如果一个试题的难度很大，但是只要被试的能力参数大于零就可认为被试掌握

了该属性吗？这不能令人信服，作者的实证研究说明不了任何问题，如果一种方法理论上站不住，所有的实证研究都失去了基础，因此，本审稿人不认为作者的研究是适当的。

回应：感谢审稿人的意见。首先，关于切点选 0 的原因原文已从两方面进行了阐述：(1)预研究和(2)在[-0.5,0.5]选取 11 个切点。从逻辑导向出发，在预研究中，当被试能力真值为连续变量时，采用 LLM 去进行参数估计，忽略掉误差后绝大都数情况下会将连续变量中的负值估计为0，而把正值估计为1，该结果与Templin和Bradshaw(2013)一文中一致。尽管Templin和 Bradshaw(2013)的主题并不是在讨论 MIRT 的认知诊断功能，但其文中仍对 MIRT 中的连续 θ 进行了切点划分(甚至是多切点划分，即将连续 θ 划分为 polytomous attribute)。其实，切点法是将连续变量转为离散变量的最常用方法，比如詹沛达和边玉芳(2015)一文中也采用了切点法对(0, 1)范围内的连续潜变量进行了切割。再比如，在结构方差模型(SEM)的测量模型中，也采用切点法将本质为连续的显变量切分为类别观察变量 Y，等等。另外，从结果导向出发，我们也探讨了共 11 个切点的效果，发现切点取 0 是一个相对最好的选择，且好到可以得到与采用 LLM 本身去分析数据一样的结果。综上所述，我们认为切点选 0 是切实可行的。

另外，审稿人所述“如果一个试题的难度很大，但是只要被试的能力参数大于零就可认为被试掌握了该属性吗？”，该问题本身违反 IRT 基本假设，即题目参数与被试参数相互独立假设。在该假设下，题目难度大或不大，与被试是否掌握某属性或被试能力值有多高并没有关系。

根据文中模拟研究结果其实已经能够说明本文欲探讨的问题。但为了体现出使用 CC-M2PLM 的优势，我们采用 CC-M2PLM、LLM 和 DINA 模型对两个实证的诊断测验数据进行了分析。根据拟合指标可知，CC-M2PLM 不仅在绝对拟合指标上拟合该数据，在相对拟合指标上也优于另外两个 CDMs。另外，我们在呈现结果部分也呈现了跨界属性（模式）和用 CDMs 估计出来的属性模式，可以看到两者存在高度一致性。

当然，本研究确实有些问题值得进一步探究，比如 MIRT 模型与 CDMs 之间的对应关系，两个类别模型参数（或量尺）之间是否均有转换关系（已有研究表明 SEM 的参数与 IRT 模型的参数可以相互转换，那么 CDM 与 MIRT 模型之间有什么关系呢？）。但需要说明的是，尽管上述问题值得今后去探究，但它们并不影响本文已做的研究对本文欲探讨的问题的回答。

第二轮

审稿人 1 意见：

意见 1：通过

回应：感谢

审稿人 2 意见：

意见 1：作者引入切点概念将能力参数转换成二分变量，并且进一步取零作为切点，只要被试的能力参数大于零就对一切属性都掌握，小于零就对一切属性都未掌握，这样的做法是不能令人接受的，是不恰当的；作者认为“审稿人所述“如果一个试题的难度很大，但是只要被试的能力参数大于零就可认为被试掌握了该属性吗？”，该问题本身违反 IRT 基本假设，即题目参数与被试参数相互独立假设。在该假设下，题目难度大或不大，与被试是否掌握某属性或被试能力值有多高并没有关系。”，不知作者为何扯上独立性的概念，建议读者读一下比较严谨的有关独立性概念的书籍后再想这个问题，如果“能力参数大于零就对一切属性都掌握，小于零就对一切属性都未掌握”真的成立，那么所有的测量问题其实都很简单了，可惜事实不是这样的；“理论上不成立的东西模拟实验却能成立”，这种想法是不具有“建设性的”，是有害的，它会把人引入歧途。审稿人深知科学研究的艰苦性，也知道作者为本文的写作下了很大功夫，然而科学研究是严肃的，不成立就不成立，这是没办法的事，审稿人并非刻意要为难作者，只是在力所能及的范围内尽一个科学研究者的良心而已。

回应：感谢您的意见。切点的选择确实是关键，如第一轮对您的回复中所述，本研究已经通过预研究和研究后再划分两种方式探讨了一个相对较为合适的切点选择方法。修改稿中，我们又从量尺角度简单阐述了下选择 0 点的合理性。也正如您担心的那样，我们相信一定存在更适合的切点选择方法，这非常值得今后进一步探讨。

另外，本研究的主要观点是已有的 MIRTM 是可以用来做认知诊断评估的，只不过因为 MIRTMs 是在连续量尺(scale)上刻画或诊断被试的潜变量值，并没有直接或明确地对被试进行分类，导致我们放大了其依子维度分排序的功能。而实际上，已有研究提及该观点(e.g., Embretson & Yang, 2013; Stout, 2007; Wang & Nydick, 2015)。我们认为由于数据本身没有发生变化，则隐藏在该数据背后的潜在建构(i.e., 潜变量)也没有发生变化，所以即便我们使用了不同的数据分析方法或模型去量化该潜变量(i.e., 刻画在不同的量尺上)，这些量化数值之

间也必定是存在某种数学转换关系的。本文选用了最易于理解的切点法来把连续量尺转换为离散量尺，该转换方法一直应用在 SEM 的测量模型之中。

我们坚持我们自己的观点，也希望可以听一下其他审稿人的意见。因此，我们已向责任编委申请将本文再送于第三位审稿专家，他/她的审稿意见如下。

最后，再次感谢您在百忙之余审阅本文，您的意见非常值得我们后续思考。

第三轮

审稿人 3 意见：

意见 1：该研究探讨的问题还是有意义的。有几点建议，请作者考虑：

切分点是本文的一个关键问题，他将被试分为掌握和未掌握两种情形。作者在本研究中的做法应该是没有问题的，而且也能保证研究的完整性。但审稿者二提出了相关问题，作者也试图进行说明，在总结部分也做了说明，既然说明了，就需要说明白究竟如何确定切分点，要么就不说。我们知道本文中的“0”是一个相对值，是在进行参数估计的过程中，将项目参数和能力参数进行了度量单位标准化后的一个人为处理结果，不同的参数估计方法或参数处理方式可能导致“0”的实际意义是不一样的。另外，能力参数是综合了被试在该维度内所有项目的作答结果估计得到的一个“综合值”，他与该维度内各项目难度参数是可以直接比较的，他们处于同一个度量单位系统上。因此，作者对审稿 2 的相关问题的回应需要重新斟酌。

回应：感谢您对本议题的认可。正如原文所述，切点是本文的关键性问题，修改稿中在讨论部分我们又添加了一些关于切点选择的思考以及为何 logistic scale 上的 0 点是一个较为合适的选择。

由于数据本身没有发生变化，则隐藏在该数据背后的潜在建构(i.e., 潜变量)也没有发生变化，所以即便我们使用了不同的数据分析方法或模型去量化该潜变量(i.e., 刻画在不同的量尺上)，这些量化数值之间也必定是存在某种数学转换关系的。比如，在二级评分题目下，基于单维 IRT 模型得到的被试能力(θ n)和基于古典测量理论(classical test theory)得到的被试真分数(Tn)存在如下转换关系：$E(X_n \mid \theta_n) = \sum_{i=1}^{I} E(Y_n \mid \theta_n) \Rightarrow T_n \mid \theta_n = \sum_{i=1}^{I} P(Y_i \mid \theta_n)$，其中 $E(X_n \mid \theta_n)$ 为能力为 θ 的被试 n 的观察总分 X 的期望，$X_n = \sum_{i=1}^{I} Y_{ni}$，Yni 为被试 n 在题目 i 上的作答，$P(Y_i \mid \theta_n)$ 为能力为 θ n 的被试 n 在题目 i 上的作答 Yni 的概率。真分数和被试能力是对同一批数据背后同一潜变量的不同量尺化结果。

另外，从量尺角度来看该问题，我们可以对 logistic 量尺上的值(θ)进行转换，将其转换到 0~1 量尺上：$\delta \equiv \frac{\exp(\theta)}{1+\exp(\theta)}$，此时 δ 所在的 0~1 量尺其实就是属性掌握概率量尺(Zhan et al., 2016; 詹沛达, 边玉芳, 2015)。通常，在该量尺下可以以 0.5 为切点将属性掌握概率转换为 0 和 1 的属性掌握状态(de la Torre & Douglas, 2004; de la Torre et al., 2010)。可发现当 δ = 0.5 时，有 θ = 0，这与预研究中得到被试潜在特质“真值”大体上以 logistic 量尺上的 0 点为分界线的结果相符。

意见 2：式 1、2、3 中，建议 p 的下标不要 1，直接说明这是正确作答概率，因为有的数据虽然是二值记分，但不是 1、0 记分。另外式子中的参数的实际意义说明还是具体一些更利于阅读理解。

回应：感谢您的建议，正确作答标记为 1 为普适性做法。描述中已添加“正确作答”。

意见 3：题目建议改为“……——以补偿模型为例”，“视角”好像认为其他视角就不是这个结果，也许这是我的语言理解问题，请作者考虑。

回应：首先感谢您的建议，经过思考，我们仍然保留原标题。尽管我们知道在 conjunctive rule 或非补偿模型下该结论是仍然适用的，但鉴于本文研究设计和结果均未涉及“补偿”规则以外的情况，因此无法使用“为例”，以“视角”为标题应是更为恰当的。为避免歧义，我们又调整了题目：

由于本文中的 MIRT 模型既是验证性的(Confirmatory)又是补偿性的(Compensatory)，所以我们将其命名为 confirmatory compensatory MIRT (CC-MIRT) model。相应的原文中的 C-M2PLM 也修改为 CC-M2PLM。同时，为更契合主题，我们修改了中文和英文标题：

中文标题：使用验证性补偿多维 IRT 模型进行认知诊断评估

英文标题：Using Confirmatory Compensatory Multidimensional IRT Models To Do Cognitive Diagnosis