《心理学报》审稿意见与作者回应

题目：多维题组效应认知诊断模型
作者：詹沛达，李晓敏，王文中，边玉芳

*由于本文在审稿阶段有细节性修改，因此不同轮的修改意见以及回应内容中可能与最终正文存在差异。

第一轮

审稿人 1 意见：文章的选题具有一定的研究和应用意义，作者将 LLM 模型中加入了题组效应，将其拓广成可进行题组效应分析的认知诊断模型，在文中作者不仅研究了新模型的返真性，也与不含有题组效应的模型进行了比较，从多方面验证新模型的有效性和可应用性。但本人认为研究还存在以下的一些问题：1、

意见 1：P7 中 C-MTECDM 中每个项目包含 K^*个项目参数，请补充说明 K^*；

回应：感谢您的提醒，我们已在文中添加说明。由于原文中 K^*已经在对 LLM 的描述时提及到了，所以我们在修改稿中对 C-MTECDM 的描述时又再次进行了强调。

意见 2：P8 中‘而“对‘最新时尚’的认知”这一潜质即可被视为一个非目标潜质(题组效应)，显然，平时更关注时尚的被试在作答该篇章阅读项目时将获得相对正收益。 ’该说明更倾向于将 DIF 效应当成是题组效应处理，与题组效应本身的定义不太一致。请作者注意题组效应本身的从含义；

回应：感谢您的提醒，考虑到可能会有部分读者出现与您类似的疑问，在修改稿中已经添加 testlet-effect(TE)与 DIF 的差异的讲解，且在原例子的基础上添加了个别修饰词。

TE 与 DIF 的相同点在于，两者都是由于题目的设定而出现了对正确作答概率 P 的影响。所不同的是，(1)TE 是由题目成组所产生的，即任一被试在回答该组题目时均可能受到 TE 的影响，且该影响(增益)对该组题目是一致的，比如某被试对题组 m 内的 10 道题目的 logit(P1/P0)均增大 γ_m，所以 TE 的参数角标 m 是表示第 m 个题目组；而 DIF 是由单一题目所产生的，其影响的是特定被试组，DIF 对正确作答概率 P 的影响(增益)对该组被试是一致的，比如第 g 组被试在回答题目 i 时的 logit(P1/P0)均增大 δ_g，所以 DIF 的参数角标 g 是表示第 g 个被试组;(2)TE 一般认为是一种随机效应，即每个被试的 γ_m 是不同的，而 DIF 一般认为是一种固定效应，即同一组内被试的 δ_g 是一样的。

意见 3：P10，表 1 中 U 矩阵的表述有混淆 U 矩阵和 Gamma 矩阵之嫌，请注意符号的统一，应该是 u1,u2,u3；

回应：答：感谢您的提醒，已修改。

意见 4：P11 中设定了题组效应参数‘3 个题组效应方差 设定为 = = =0.5。’但如何生成相应的 Gamma 参数呢？另，参数估计时，文中并没有报告 Gamma 参数的估计精度，其标准差估计的好坏不能代替 Gamma 参数的估计好坏。而 Gamma 参数本身是一种被试参数，其值的大小反应了被试多在程度上需要作用属性之外的东西进行作答，直接影响作答概率，所以该参数的结果是需要报告的。

回应：感谢您的建议。首先，该问题涉及到 TE 参数的操作定义或量化定义，即(以往关于 TE 的所有研究均)设定 TE 满足平均数为 0，方差为 σ^2 的正态分布，且 TE 的大小是由其方差为 σ^2 的大小决定。原文已经在引言和模型建构部分多次提及 TE 的量化定义。因此，在参

数生成部分直接给出了各 TE 方差的大小。由于 TE 的大小是由其方差的大小表示的，因此仅需报告 TE 的方差即可，这与以往所有关于 TE 的研究一致。

其次，根据参数估计方法可知，通过了解其他参数的返真性即可推知单一 γ 参数估计的好坏。因此，在上千样本量的情况下，通常并不需要列出一巨大的表格去说明某潜质参数的估计结果表格，而仅需列出一相对小的项目参数估计结果表格。亦可参见其他关于 MIRM 的研究。所以，我们并未采纳该意见。

意见 5：作者文中使用的 Bias 指标，但本人认为该指标可能不太适用，因为该指标反应的是某一参数在多次实验下的平均偏差，由于偏差具有符号，并不能很好的反映参数估计的返真性。如果想使用该指标，最好报告该参数在多次实验下的最小值，最大值，平均数以及标准差或方差。前两者体现返真性的好坏，后者体现估计的稳定性，是反映一个模型好坏及可用性的指标。建议补充模型稳定性的相关信息。

回应：感谢您的建议。首先，Bias 指标是用于反映参数估计返真性中估计偏差的指标，而用于反映参数估计精确性的是 RMSE 指标。其次，为更好地体现参数估计结果，我们采纳了您的建议，在修改稿中添加了用于反映参数估计是否可被接受的 ARB 指标和反映参数估计稳定性的 SD 指标。

审稿人 2 意见：

意见1：“2 多维题组效应认知诊断模型的开发”部分第2段第1行，“项目内单维题组效应”容易理解。但是对于“项目内多维题组效应”，即某个题目对应多个题组效应，这种情况在实际测验中存在或常见吗？

回应：相比于项目内单维题组效应，项目内多维题组效应要更难理解这是符合思维逻辑的。但实际上，项目内多维题组效应比项目内单维题组效应更为常见，只是在以往研究中被忽略了。由于篇幅原因，原文中已给出该概念的参考文献供不理解的读者进行查阅：詹沛达, 王文中, 王立君, 李晓敏. (2014). 多维题组效应Rasch模型. *心理学报, 46*(8), 1208–1222。

传统的“题组”是指最简单的捆绑式题组，比如阅读理解。而根据詹沛达等(2014)的研究可知，“捆绑在一起的项目”并不等同于“题组”，而没有捆绑在一起的项目也并不是说它们就不可能是题组，比如，采用同一题型的所有项目、考查同一单元内容的所有项目，它们可能分布在同一test的不同位置，但它们的组合是会对被试的作答产生影响的，因为有的被试倾向于做选择题，而有的被试倾向于做填空题等等。

此外，从模型参数角度讲，多维题组效应向量也比单维题组效应更具普适性。因此，本研究直接使用了多维题组效应向量，以期提高模型的普适性。

意见2：公式 (3) 下面1段第3行中，什么是“分层模型”？

回应：感谢您的提醒，修改稿中已添加相对应的英文，并将“分层模型”修改为“层级项目反应模型(hierarchical IRM)”。

意见3：公式 (5) 下面第6行 “λ_{ik} 为项目 i 中属性 k 的权重(即 $\lambda_{ik} > 0$)，用于描述掌握属性 k 对正确作答项目 i 的概率的增量”，但是在DINO模型中，至少掌握1个项目考查属性时的正确作答概率与掌握多个项目考察属性时的正确作答概率相等，此处 $\lambda_{ik} > 0$ 的约束，有不合理之嫌；

回应：感谢您的提醒，原文中就已经说明了补偿、非补偿（连接）和分离之间的关系和区别。

首先，本文仅探讨了Logistic题组框架(LTF)中补偿和非补偿两种情况下的模型，而DINO模型为分离模型，并不在本文探讨的范围之内。其次，在补偿模型中，属性权重大于0是必须的且符合思维逻辑的。再次，考虑到可能会有部分读者出现与您类似的疑问，修改稿已经

在描述LTF部分添加了如何将LTF转化为分离模型的方法（$h(\cdot)=\lambda_{i(K)}[1-\prod_{k=1}^{K}(1-\alpha_{nk})^{q_{ik}}]$，同样$\lambda_{i(K)}>0$），恕于作者精力和篇幅有限且避免赘述，并未进行详细探讨。

意见4：“2.4 多维题组效应认知诊断模型框架”部分第1段提出“LTF (Logistic testlet framework) 可描述为：$\log(P_{ni1}/P_{ni0})=intercept+target\ latent\ traits+testlet\ effects$ ”。但是在实际情形中，题组与目标潜在特质之间可能会存在交互效应，而文中认为题组效应对于正确作答仅仅是一种补偿效应，有不合理之嫌；

回应：感谢您的审阅，但我们认为该问题缺乏专业性。已有文献中的常见做法是主张题组效应与目标潜在特质之间不存在交互作用(Bradlow et al., 1999; Wainer et al., 2000; DeMars, 2006; Wang & Wilson, 2005; Huang & Wang, 2013; 詹沛达, 2014)，因此在LTF中模块1与模块2是求和(补偿)关系。另外，相关文献已有探讨，认为题组效应等价于高阶反应模型中的误差项，而假设误差项与目标项之间为独立关系是自CTT到IRT的一贯做法。

意见5：提出一种新的模型，除了提供模型的公式呈现以及参数说明，还需要详细提供对新模型参数进行估计的参数估计思路与原理；

回应：感谢您的建议，我们已在修改稿中添加“3 参数估计”一章节。

意见6：研究一表2中，对于C-MTECDM模型，有多达8个题目 (题目5、13、19、21、24、25、28以及30) 的截距参数的Bias值大于0.1 (偏大)，还有多个题目的Bias值接近0.1；对于N-MTECDM模型，第10个题目的截距参数的Bias值也大于0.1。这确实不能说“两个模型对题组效应参数方差和截距参数的返真性均较好”。同理，在表3中，对于C-MTECDM模型，也有部分题目的属性权重参数的Bias值或其绝对值较大，如第5题的λ_{i5}；对于N-MTECDM模型，多个题目的$\lambda_{i(5)}$的Bias绝对值都偏大。这说明对各属性权重参数的返真性并不是特别好。研究二的结果也存在类似情况。

回应：感谢您的建议，考虑到MTECDM的复杂性(同时包含类别变量潜质和连续变量潜质)和**Q**矩阵的设定(共包含8个维度潜质)，根据已有关于MIRM的研究可推断出欲对MTECDM实现较精准参数估计很可能需要较大的样本量和重复次数。由于原稿中，样本量设定和重复次数的原因，导致参数估计结果的展现并不稳定好看。

为了给予读者一个在使用新模型时关于样本量的指导性建议，我们在修改稿中我们设定3个样本容量(1600、3200和4800)以期探究MTECDM的参数估计返真性，新的参数估计结果显示在3200以上的样本量参数估计结果很好。并在结果中指出：C-MTECDM的建议样本量应大于3000，而N-MTECDM的建议样本量应大于1500

意见 7：“引言”部分第 1 段第 1 行及第 6 行，认知诊断评估与认知诊断模型的对应英文有误，应改为“cognitive diagnostic assessment”与“cognitive diagnostic model”

回应：两种用法具有研究者使用，尊重审稿人的意见，修正稿已经修改。

意见 8：“引言”部分最后 1 段第 3 行存在错误字，“即”应改为“既”；

回应：感谢您的细致审阅，但根据汉语正规用法，表示“也就是”的是“即”字。

意见 9：“2 多维题组效应认知诊断模型的开发”部分第 1 段第 2 行，什么是“测验目标潜质”？这里的“潜质”是指“潜在特质”吗？

回应：根据简约原则，通常把“潜在特质”简写为“潜质”。另外，本文将 latent class 与 latent trait 均称为潜质。

“测验目标潜质”根据其字面意思即可理解，是某测验所真正要测量的潜质，即测验编

制时所规定的测验目标。我们认为这是文字理解问题，因此并未修改。

意见 10：“2 多维题组效应认知诊断模型的开发”第 3 段，由于 $\mathbf{u}'_{im}\boldsymbol{\gamma}_{nm}=u_{i1}\gamma_{n1}+\cdots+u_{im}\gamma_{nm}+\cdots+u_{iM}\gamma_{nM}$ 且 $\boldsymbol{\gamma}_{nm}=(\boldsymbol{\gamma}_{n1},\boldsymbol{\gamma}_{n2},...,\boldsymbol{\gamma}_{nM})'$，所以等式 (2) 存在两个小错误：① $\mathbf{U}_{im}$ 应该是 $\mathbf{U}'_{im}$；② 等式 (2) 的等号右边中行代表题目，所以 $\mathbf{U}_{im}$ 不应该包含下标 i；

回应：感谢您的细致审阅，我们已经将 $\mathbf{U}_{im}$ 修改为 $\mathbf{U}'_{im}$。而等式 (2) 的等号右边并未说明行代表题目，所以 $\mathbf{U}'_{im}$ 仍保留了下标 i

意见 11：“2.2 模型建构基础——LLM 简介”部分第 1 段第 2 行，I 与 K 分别代表什么，文中没有进行说明。同一行中，“定界”应改为“界定”；
回应：感谢您的细致审阅，我们已经添加说明，并进行相应修改。

意见 12：公式 (6) 下面第 4 行“……抽离出由非目标潜质的题组效应”表述不清
回应：感谢您的建议，我们已经修改。

意见 13：“2.4 多维题组效应认知诊断模型框架”部分第 1 段第 3 行，“…… Rasch 题组模型 (见式(2))”有误，应改为“…… Rasch 题组模型 (见式(3))”；
回应：感谢您的细致审阅，我们已经修改。

意见 14：公式 (19) 下面 1 行应该对公式(17)、(18) 与 (19) 中的 n_{kr} 及 n_{pr} 等符号进行说明，而不是对 n_k 以及 n_p 进行说明；

回应：感谢您的建议，我们已经修改。

意见 15：公式 (22) 下面 1 行，L 表示“似然函数”，而不是“极大似然函数”；
回应：感谢您的建议，我们已经修改。

意见 16：“5 总结与展望”部分第 2 段第 1 行，“…… 2.3GHz 的 Inter Core i5-2410M 处理器……”应改为“…… 2.3GHz 的 Intel Core i5-2410M 处理器……”；
回应：感谢您的建议，我们已经修改。

意见 17：参考文献格式错误，“Rupp, A. A., Templin, J., & Henson, R. A. (2012). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.”应为“Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.”
回应：感谢您的细心审阅，但此处并没有错。
*修改稿在精简全文时已经删除本条参考文献。

审稿人 3 意见：本文拟要解决的问题在当前教育与心理测量领域中是值得关注的。但文章存在如
下问题：作者在摘要中说到： “但目前仍未有将两者进行结合的先例。 ”根据本审稿人掌握的资料，University of California, Los Angeles 的 Hansen, M & Cai, L(2013)在International Meeting of the Psychometric Society(IMPS)会议上已经发表过同样的研究了，……本审稿人支持作者的创新精神，但希望作者今后应该及时掌握国际上的最新研究结果，避免再次发生这种情况。鉴于同样的研究已存在，审稿人建议退稿。希望作者不要灰心，可以参考相关研究将该问题做得更加深入一些。

回应：(1)Hansen & Cai (2013)为会议论文，我们这里将探讨 Hansen (2013)这篇更为详细的博士论文；

(2)由于本文与 Hansen(2013)一文的建构逻辑不同，在正文中进行引用恐会影响本文的流畅性。综合考虑后，我们在修改稿的讨论部分添加了一小部分内容，用于简单描述本研究和 Hansen 的研究的差异。限于正文篇幅，这里作出详细解释：

关于该议题的研究目前有刘文(2011)和 Hansen(2013)两篇博士论文涉及，刘文(2011)和 Hansen(2013)从不同的角度分别提出了题组认知诊断模型(testlet CDM, TCDM)和层级诊断模型(hierarchical diagnostic model, HDM)。其中，TCDM 是基于三参数题组模型(Wainer et al., 2000)所建构的，两者区别在于描述目标潜质时，TCDM 假设目标潜质为类别变量，而三参数题组模型假设目标潜质为连续变量。TCDM 的项目反应函数可描述为：

$$P_{ni1}=c_i+\frac{1-c_i}{1+\exp[-(\tau_i+\sum_{k=1}^{K}(\beta_{ik}\alpha_{nk}q_{ik})-\gamma_{nd(i)})]}\quad , \tag{24}$$

式中，

c_i为题目i的猜测参数;

τ_i为题目i的截距；

β_{ik}为属性k在题目i上的斜率；

$\gamma_{id(n)}$为第n个被试题组d的题组效应；

α_{nk}与q_{ik}含义同上。显然，TCDM属于补偿模型，与同属补偿模型的C-MTECDM对比后可发现2点不同：(1)TCDM包含一个猜测参数c，但原文在研究中直接将该参数设定为了0，并未进行探讨。我们认为在CDA中，为了与猜测参数g相区别，该参数应根据其实际意义被称为下渐近线参数。而在IRM中，下渐近线参数是用于描述当潜在特质趋向于-∞时正确作答概率的最低值(项目特征曲线的下渐近值)，但对于CDM，其正确作答概率的最低值一定是在认知属性模式$\boldsymbol{\alpha}=\mathbf{0}$处，即正确作答概率不存在渐近值，反之，CDM中也就不应存在参数c。这也是目前所有Logistic形式下的CDM并不含下渐近线参数的原因；（2）$\gamma_{nd(i)}$是项目内单维题组效应参数，无法处理可能存在的项目内多维题组效应。此外，该研究并未探讨模型中各项目参数的返真性。综上所述，虽然TCDM的建构并不严谨，但它是对解决CDA中可能存在的题组效应这一问题的首次尝试，对后续研究起到了“抛砖引玉”的效果。

HDM是基于双层项目因素分析模型(2-tier item factor analysis model) (Cai, 2010)建构的，

两者区别是在主层(primary tier)中，HDM假设潜质为类别变量，而双层项目因素分析模型假设潜质是连续变量。HDM的项目反应函数可描述为：

$$P_{ni1}=\frac{1}{1+\exp[-(\alpha_i+h(\boldsymbol{\gamma}_i,\mathbf{q}_i,\mathbf{x})+\beta_{is}\zeta_s)]} \quad , \tag{25}$$

式中，

α_i为题目i的截距参数；

$h(\boldsymbol{\gamma}_i,\mathbf{q}_i,\mathbf{x})$定义了认知属性纳入项目反应函数的形式；

ζ_s用于描述第s个题组效应；

β_{is}是项目i在题组效应ζ_s上的斜率参数。

在HDM中，$h(\boldsymbol{\gamma}_i,\mathbf{q}_i,\mathbf{x})$的功能与LTF中模块1的功能类似(见正文)，即HDM和MTECDM均可在CDA情境下根据测验情境在补偿模型、非补偿模型和非连接模型等之间进行转换。两者不同处在于$h(\boldsymbol{\gamma}_i,\mathbf{q}_i,\mathbf{x})$仅限于CDA情景；而LTF中模块1更多地是一个概念性模块，它完全可由测验分析人员进行自定义，不仅适用于潜质为类别变量的诊断测验还适用于潜质为连续变量的传统IRT测验(见正文)，即LTF灵活性相对更大。此外，对比HDM和本文提出的MTECDM后可发现两者的主要区别还在于对题组效应的描述上，(1)由于flexMIRT (Cai, 2012)的限制(或者说是双层项目因素分析模型的限制)，HDM要求每个题目最多只能归入1个题组或特殊组维度(group-specific dimension)之中(Hansen, 2013)，也即ζ_s为项目内单维题组效应参数，无法处理可能存在的项目内多维题组效应；(2)HDM中β_{is}是斜率参数，用于描述题组效应ζ_s对正确作答概率的贡献的加权。而MTECDM中$\mathbf{u}_{im}$为指标向量，用于指示被试n对题组项目i的反应是否受到第m个题组效应的影响，而对正确作答概率的贡献完全由题组效应大小决定。另外，在Hansen(2013)的研究中并未探讨当使用不含题组效应的模型去分析含有题组效应的CDA数据时会给分析结果带来什么危害，而本研究二中已对该问题进行了较为详细的探讨。综上所述，虽然HDM较TCDM来说更为严谨、灵活，但由于flexMIRT的限制，HDM仍无法处理可能存在的项目内多维题组效应。

HDM、TCDM和MTECDM三者间之所以存在一定程度的相似性这与它们的建构过程和建构基础有关：(1)HDM是对双层项目因素分析模型的修改，而双层项目因素分析模型又是对双因素模型(bi-factor model) (Gibbons & Hedeker, 1992)和MIRM的综合拓广；(2)TCDM是对(三参数)题组模型的修改，而通常认为项目内单维题组效应模型又是双因素模型的约束模型(Li et al., 2006; Wainer et al., 2007; DeMars, 2006, 2012; 詹沛达, 2014)；(3)MTECDM是对LLM和项目内多维题组效应模型的综合拓广，而LLM又是对(单层)项目因素分析模型的修改且多维题组效应模型是对单维题组效应模型的拓广，同时项目内多维题组效应模型与层级项目反应模型之间又具有近似等价性(詹沛达, 2014)，而当层级项目反应模型仅包含2层因素时就是双因素模型。可以看出，HDM、TCDM和MTECDM的建构基础自身就具有一定关联和相似性，因此，当研究者从不同的出发点去推敲如何解决同一问题的方法时存在某种程度的

思路交叉也是符合逻辑的，比如IRT领域最经典的思路交叉例子是Rasch模型和单参数IRM。总之，除了TCDM中不合理的*c*参数外，且由于项目内单维题组效应参数是项目内多维题组效应向量的特例，所以TCDM和HDM亦可归属于LTF之中。

第二轮

审稿人 1 意见：

谢谢作者耐心的回复，以及所做的修改和补充。该研究具有创新性，虽然个人比较认同其中一位专家增加一个应用研究的建议，但鉴于实际情况，本文的研究量已经足够大了。对于本文关于 bias 指标部分的变化，个人倾向你保留图的形式，这样更加直观有效。

回应：感谢审稿人对本文的肯定，根据我们的理解审稿人应该是笔误将“表”写成了“图”。综合考虑后，我们将研究一中 N-MTECDM 的报告结果还原为了表格形式，这样更加直观有效。

审稿人 2 意见：

意见 1：本文将已有的线性逻辑斯蒂克模型 (LLM) 和多维题组效应 Rasch 模型 (MTERM) 两者进行结合，从而实现“既能进行认知诊断又能处理题组效应”的功能，创新性及理论价值不高。

回应：感谢审稿人对本研究的评论。有时候一个好的 idea 恰恰看似那么简单，创新性和理论意义并不等同于一堆繁杂的数学推导，其本质应在于是否能够有效地解决一个目前少有研究者探究的实际问题。

本研究除了提出 2 个特殊模型外，更主要的是提出了 Logistic 题组框架。本研究的主要目的在于探索出一种可处理 CDA 中可能存在的题组效应的方法。全文书写逻辑是“从特殊到一般”，具体是指：先从读者常见的“特殊 CDM(i.e., DINA 和 LLM)”入手，把它们与目前处理题组效应最普适的 MTERM 相结合，提出了“特殊的”C-MTECDM 和 N-MTECDM，实现了本研究的目的，即“既能进行认知诊断又能处理题组效应”。这样做的好处是“特殊模型”易于读者理解。

之后把这两个“特殊的”MTECDM 与目前已有的题组反应模型综合起来，提出了更“一般的”logitstic 题组框架，该框架几乎囊括了目前所有的（无论是 IRT 范围内的还是 CDA 范围内的）题组模型。这样做的好处是，有助于读者更深入地理解各个题组模型的本质共同点。

当然，我们的方法也并非十全十美，其中存在的不足我们已经在文中讨论部分进行了简单说明。整体来看，我们认为本研究仍有相当的创新性和理论价值。

意见 2：Reviewer 2 对作者在第 4 条审稿意见上的回答不满意。个人并不认同作者所说的“……该问题缺乏专业性”，理由如下：(1) 以往研究都是假设题组效应与目标潜在特质之间不存在交互作用，并不代表两者本质上就不存在交互作用；(2) 作者提到“相关文献已有探讨，认为题组效应等价于高阶反应模型中的误差项……”，请给出是哪篇文献，而且请注意本文并不是探讨高阶反应模型。

回应：我们理解审稿人的意思，但同时认为该意见已经偏离本研究所探讨的范围。之前我们说审稿人对该问题的建议缺乏专业性是特指 testlet 领域内的知识，而非广义的统计和测量范围内的知识。原因如下：（1）如之前所述，在已有的所有题组模型中均假设题组效应和目标潜变量之间为正交关系，其主要原因是便于统计和意义解释。因此，我们也默认使用了该假设；（2）除题组模型外，bi-factor models 也同样假设主维度和子维度之间为正交关系。在 IRT 范围内，已有研究探讨了 bi-factor models, testlet response models 和 higher-order models 三者之间关系，见：

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145–168.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*(2), 104–121.

Li, Y. M., Bolt, D. M., & Fu, J. B. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3–21.

Rijmen, F. (2009). *Three multidimensional models for testlet-based tests: formal relations and an empiric comparison* (Research Report. No. RR-09-37). Princeton, NJ: Educational Testing Service.

Huang, H.-Y., Wang, W.-C., Chen, P.-H., & Su, C.-M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement. 37*(8), 619-637.
詹沛达. (2014). *多维题组效应模型的开发与应用.* 硕士学位论文, 浙江师范大学.

下面附上一个 IRT 范围内的简单推导，如下：

$$
\begin{aligned}
&\theta_{pt} = a_{pt}\theta_p + \varepsilon_{pt} \\
&g(\pi_i) = a_{ti}(\theta_{pt} - \xi_{ij}) && \Leftarrow \text{二阶模型} \\
&\quad = a_{ti}\theta_{pt} - \delta_{ij} \\
&\quad = a_{ti}a_{pt}\theta_p + a_{ti}\varepsilon_{pt} - \delta_{ij} \\
&\quad = a_{pi}\theta_p + a_{ti}\theta_{pt} - \delta_{ij} && \Leftarrow \text{双因子模型} \\
&\quad = a_{pi}(\theta_p + \theta_{pt}/a_{pt}) - \delta_{ij} \\
&\quad = a_{pi}(\theta_p + C^*_{pt}\theta_{pt}) - \delta_{ij} \\
&\quad = a_{pi}(\theta_p + \gamma_{pt} - b_{ij}) && \Leftarrow \text{TRM}
\end{aligned}
$$

(引自詹沛达. (2014). *多维题组效应模型的开发与应用.* 硕士学位论文, 浙江师范大学.)
可见 $\varepsilon_{pt}=\theta_{pt}=a_{pt}*\gamma_{pt}$，。而假设误差项与目标潜变量之间相互独立是目前绝大多数统计测量方法的基本假设。当然，我们把 IRM 范围内的结论直接使用在 CDM 中可能存在的不足也已经在正文讨论部分提到了；（3）“假设”一定是存在局限性的，所以如果审稿人非要让我们去探讨当误差项与目标潜质见相互独立这一假设不能满足时会对分析结果带来什么影响，恕于精力、篇幅和研究主题等原因，我们暂无法做到，希望这会成为我们今后的一个研究点之一。

意见 3：“1 引言”部分最后 1 句话呈现的脚注 1 部分，个人觉得标注内容不妥。Hansen (2013) 的博士论文或 Hansen 和 Cai (2013) 的会议论文虽然并未发表，但 Reviewer 3 在审稿意见中有提到 Hansen 和 Cai 的已有相关研究，严谨、专业且谦虚的做法应该是：引用他们的工作并且在稿件中简单描述本文与他们研究的差异 (个人认为这部分内容的呈现比字数的限制更重要)，然后再进行标注。
回应：感谢您的建议。由于两者建构角度不同，在正文中进行引用会影响本文的流畅性。我们在修改稿的讨论部分添加了一小部分内容，用于简单描述本研究和 Hansen 的研究的差异。

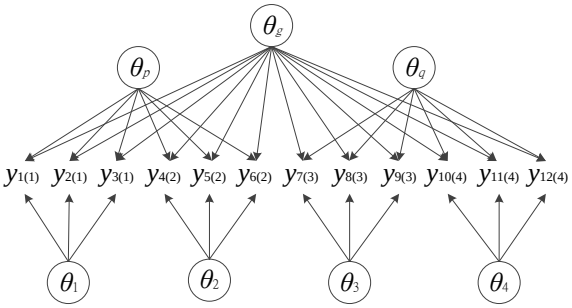
意见 4：个人认为，图 1 (a) 描述的题组效应应该是项目间多维题组效应，理由有二：(1) 存在多个题组效应；(2) 每个题目仅对应一个题组效应。
回应：您说的对，“项目间多维”和“项目内单维”本质是一样的，只不过描述的角度不同。而它们均可以看成是“项目内多维”的特例，“项目内多维”和“项目间多维”是 MIRM 中的基本概念。

意见 5：“3 参数估计”部分存在的问题：(1) 没有给出似然函数 $L(\boldsymbol{\lambda};\boldsymbol{\alpha},\boldsymbol{\gamma})$ 的表达式；(2) $\boldsymbol{\alpha}$ 初值的设定方法也没有给出。在 MCMC 算法中，$\boldsymbol{\alpha}$ 的初值非常重要，因为如果 $\boldsymbol{\alpha}$ 的初值与真实的 $\boldsymbol{\alpha}$ 的初值接近，MC 链可以快速地收敛于平稳分布；如果 $\boldsymbol{\alpha}$ 的初值设置得不好，要达到收敛可能需要更多的时间。
回应：感谢您的审稿，限于篇幅原因，根据条件独立性假设或广义局部独立性假设，似然函数读者可自行推断，认知属性的初值均按 *Bernoulli*(0.5)随机生成。初始值的设定的确会会影响参数估计值，但收敛时间暂不是本研究要考虑的问题。

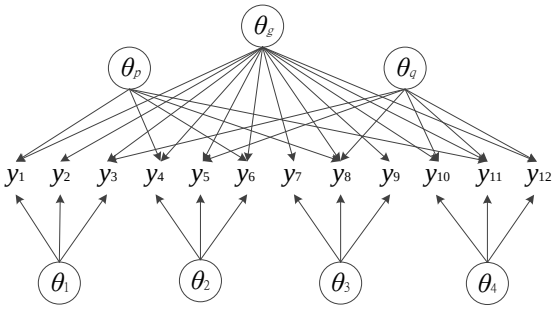
意见 6：“4.1 研究设计”部分存在几个问题：(1) 按照表 1 的方式设定 **Q** 矩阵与 **U** 矩阵有什么依据或者考虑吗？(2)“4.1.2 项目参数、题组效应与被试属性掌握模式的设定”部分中，并没有看到

被试属性掌握模式是如何设定的；(3) 公式 (25) 的描述有误，因为该指标与属性 k 有关，所以应该含有下标 k。

回应：感谢您的审稿。由于本研究关注的重点并不在 **Q** 矩阵这里，所以 **Q** 矩阵设定在已有研究的基础上相对“simply and casually”；**U** 矩阵的设定是为了体现出多维题组效应参数的优势，而设计了交叉分类情况，即同一个题目可能同时涉及多个题组效应（子维度），如下图所示：



(a)不存在交叉分类结构



(b)存在交叉分类结构

(引自詹沛达. (2014). *多维题组效应模型的开发与应用*. 硕士学位论文, 浙江师范大学.)

对于被试属性掌握模式的设定我们之前在缩减篇幅时不小心删除掉了，现已复原至修改稿中，具体设定方法为：目标属性数 K=4，即共有 2^4=16 种属性模式，设定每种属性模式人数分别为 100、200 和 300 人，则共 1600、3200 和 4800 人。另外公式(25)已做出相应修改。

意见 7：“5.3 研究二结果与结论”部分存在的问题：表 4 中用 LLM 模拟的数据，用 LLM 模型去拟合反而不如用 C-MTECDM 拟合得好。这个如何解释？

回应：首先从结果上解释，AIC 和 BIC 在十几万的基数下相差几百或几十，可以认为两者是“没有差异的”，而实际分析出的数据中的差异很可能是由于随机误差导致的。

其次从逻辑上解释，这也是 nested model 研究中会出现的问题，用 nested model 作为 true model 去生成数据时候，由于随机性的存在（e.g., 随机生成一个 r 与正确作答概率 p 比大小去判断该题的得分），nested model 很有可能在拟合度上略低于 general model，这是因为 general model 的参数更多，其中某些参数会把一部分模拟作答中潜在的误差给抽离出来，尤其是本研究中题组效应参数正是为了处理题组效应这一误差项而添加的。所以导致拟合度的微弱提高很可能是一个副产品。

意见 8：第 10 页倒数第 2 行，公式是由一个列向量的转置乘以一个列向量而得到，怎么可以说是由两个矩阵组成呢？

回应：感谢您的意见，从公式讲是列向量转置和列向量的乘积得到的，从逻辑上讲也可以说

是由两个矩阵的组成。

意见 9：公式 (2) 中，i 和 m 是矩阵 **U**'的下标吗？如果是的话，$\mathbf{U}'_{im}$ 是一个矩阵，那么 **U**'代表什么？如果没有理解错的话，应该不含下标。

回应：您的理解是正确的，我们已做了修改。

意见 10：公式 (2) 下面一行"当 $\mathbf{U}'_i$ 中只有 1 个元素为 1 时……"的描述不准确。要由 (1) 式得到 (3) 式，必须进行更为准确的描述，比如，"$\mathbf{U}'_i$ 的第 m 个元素为 1"。

回应：感谢您的建议，我们已做了修改。

意见 11：公式 (4) 表述有误，应该描述为 **Q** 而不是 $\mathbf{Q}_{ik}$，也不应包含下标。

回应：感谢您的建议，我们已做了修改。

意见 12：公式 (4) 下面第 6 行"则 LLM 将被试在项目 1 上的属性掌握模式"的表述不准确，改为"在 LLM 中项目 1 将被试的属性掌握模式分为 4 组"可能更好。以下类同。

回应：感谢您的建议，但我们认为该句修改为"在项目 1 中 LLM 将被试的属性掌握模式划分为 4 组"。

意见 13：对公式 (5) 中的符号进行描述时，建议也加上对 P_{ni0} 的解释。

回应：感谢您的建议，我们已做了修改。

意见 14：公式 (5) 下面第 5 行中，准确地讲，λ_{i0} 应该是"正确作答项目 i 的概率的对数发生比的基线"。$\exp(\lambda_{i0})/[1+\exp(\lambda_{i0})]$描述的才是正确作答项目 i 的基线概率；公式 (5) 下面第 6 行中，准确地讲，λ_{ik} 应该是"掌握属性 k 对正确作答项目 i 的概率的对数发生比的增量"。

回应：感谢您的建议，我们已做了修改。

意见 15：第 12 页倒数 2 行，准确地讲，应该是"……对正确作答概率的对数发生比有补偿作用的 CDM"。

回应：感谢您的建议，我们认为两种描述方式所表达的内容本质是一样的。所以，做了如下修改"假设各认知属性对正确作答概率(的对数发生比)有补偿作用的 CDM"

意见 16：公式 (6) 下面第 4 行和第 8 行中，符号 γ_m 与 $\boldsymbol{\gamma}_m$ 的使用有问题；公式 (6) 下面第 9 中的"对角阵 **Σ** 为"改为"协方差矩阵 **Σ** 为对角阵"可能更好。

回应：感谢您的建议，我们已做了修改。

意见 17：公式 (9) 下面第 1 行和第 2 行中"……对正确作答项目 i 的概率的增量"表述不正确，应该是"……对正确作答项目 i 的概率的对数发生比的增量"。

回应：感谢您的建议，我们已做了修改。

意见 18：如果能够在图 2 中每种模型旁边标上模型名称的话，将会更清楚、更准确。而且，图中多处 $\boldsymbol{\gamma}_m$ 应该改为 $\boldsymbol{\gamma}_n$，意为被试 n 的题组效应 (与上下文保持一致)。

回应：感谢您的建议，我们已做了修改。

意见 19：公式 (13) 下面第 6 行，$h(\cdot)=\sum_{k=1}^{K}\lambda_{ik}\alpha_{nk}q_{ik}$ 应改为 $h_n(\cdot)=\sum_{k=1}^{K}\lambda_{ik}\alpha_{nk}q_{ik}$，以保持上下文的一致。以下类同。

回应：感谢您的建议，我们已做了修改。

意见 20：公式 (14) 上面第 2 行的公式"γ_{nm}~$MVN(\mathbf{0},\boldsymbol{\Sigma})$"表述有误，根据对上下文的理解，$\gamma_{nm}$ 是单个变量，它怎么会服从多变量正态分布呢？

回应：感谢您的建议，我们已做了修改。

意见 21：公式 (20) 上面第 4 行中的"MIRM"是什么的缩写？是指多维项目反应模型吗？

回应：是的。

意见 22："5.3 研究二结果与结论"部分第 1 段第 1 行，"其中，表 5……"应该改为"其中，表 4……"。

回应：感谢您的建议，我们已做了修改。

责编复审意见：

a.Word counts alright, now 9000+ words,

b. Reference is 30, alright,

c. The English abstract is pretty alright, I have done extremely minor changes.

回应：感谢您对本文的肯定，我们已经按照您的建议对英文摘要做了进一步修改。另外综合考虑本次外审意见，我们对文章进行了小部分修改（正文用棕色字体表示修改内容）。