《心理学报》审稿意见与作者回应

题目：一种基于多阶认知诊断模型测评科学素养的方法
作者：詹沛达，于照辉，李菲茗，王立君

第一轮

审稿人 1 意见：

该文从“为学习而评价(assessment for learning)”的新测评理念出发，试图解决如何基于大规模测评数据 PISA2015，实现对科学素养的准确测评的问题。文章尝试从认知诊断评估的视角来探究如何测评科学素养，并提出新的理论模型来解决该问题，这是一个有益的尝试。作者从实际问题出发，寻找解决的途径，建立了多阶认知诊断模型，而且提出了模型参数估计方法，进行了模拟研究。文章有一定的理论深度。

意见 1：

文章希望“实现对科学素养的准确测评”，由于是认知诊断，评估“准确测评”的标准本来是将分析的结果和（某一些随机抽取）被试的实际状况进行对比，考察吻合程度，当然对于这一批数据是无法实现的。模拟研究似乎可以部分弥补这一缺陷（因为实际数据和模拟数据是有差别的），但是模拟结果的报告（图 9），不知道为什么不报告模式判准率，而仅仅报告属性判准率？如果模式判准率低，那么说“准确测评”就有一点勉强。

回应：

感谢您的意见，我们也赞同您的观点。原文中，为提高图片呈现效果，图 9 仅呈现了 ACCR。实际上 PCCR 已在原文中进行说明：“另外，PCCR 为 0.512，考虑到属性数量为 9，即有 512 种可能的属性模式需要被估计，根据已有研究经验，该判准率符合预期。”，请见正文第 12 页。

意见 2：

作者使用认知诊断手段对科学素养的 12 个第一阶子属性进行分析。这 12 个第一阶子属性的“粒度”都比较大，也就是说，都不是“细粒度”。有的学者认为认知诊断需要针对“精细粒度”（例如可以参见 Leighton & Gierl[2007]所编著的教育认知诊断评估，理论和应用一书中最后一篇文章 Gierl & Leighton,2007,p.346）。请问作者如何考虑这个问题。

回应：

感谢您的意见。我们认为您的疑问实际上是认为 CDM 更适宜于测量含义或粒度较小的潜变量，而本文中的第一阶子属性的粒度较大，或许使用 MIRT 模型来进行分析可能更好。

首先，我们不反对您以及 Leighton 等研究者的观点，如詹沛达，陈平和边玉芳(2016)所述:“MIRTMs 与 CDMs 的主要区别在于前者假设潜变量空间由 K 维连续变量(即潜在特质)组成，而后者假设潜变量空间由 K 维离散变量(即属性)组成。而至于如何在这两类模型中进行选择，则通常取决于测验目标潜变量的含义和测验目的，若目标潜变量的含义相对较大或粒度较大(比如数学能力)则 MIRTMs 更为常用，而若含义相对较小或粒度较小(比如分数进位)则 CDMs 更为常用;若测验目的是探究被试在某方面能力的大小则 MIRTMs 更为常用，而若测验目的仅为诊断被试是否掌握某方面的能力则 CDMs 更为常用。当然,万事非绝对,暂无研究和证据表明我们一定要按某种规则选用两者中的某一个。”因此，我们认为要依据测验分析者的目的来选择 CDM 还是 MIRT 模型（甚至是 CFA）。本研究的目的是欲在 CDA 中实现对科学素养的测评，因此，选择 CDM 做基本的分析模型是符合研究目的的。

其次，在修改稿的讨论部分，我们增加了一段针对该问题的内容，请见正文讨论部分。

意见 3：

文章正文 p.5,Line12,“被试对各第一阶属性的掌握满足条件独立”，这句话的准确含义是什么？看上去有一点吃力。

回应：

条件独立性假设应该是所有潜变量模型的基本假设，通常，它假设给定或固定潜变量后，受该潜变量影响的各（观察）变量之间相互独立，即假设各（观察）变量之间的相关关系完全由给定的潜变量所解释。譬如，IRT 模型中的条件独立假设是指假设给定潜在能力后，被试对各个题目的作答（概率）之间相互独立。类似地，在高阶潜在模型中，条件独立假设是指假设给定高阶潜在特质后，被试对各属性的掌握（概率）之间相互独立。

为便于读者理解，修改稿中我们调整了该句的位置，并进行重述，请见正文第 5 页。

意见 4：

文章 3.2.3 节说“超先验(hyper prior)分布设定为：$\mu_\beta \sim N(-1.096, 4)$, $\mu_\delta \sim N(-1.096, 4)$,”，请问为什么是同分布？该如何解释？

回应：

感谢您的问题。由于 $\text{logit}(-1.096) \approx 0.25$，所以设定 β 和 δ 的均值的超先验分布为均值为–1.096 的正态分布，是与四选一题目的理论猜测概率相符合的。另外，因为我们设定该超先验分布的方差为 4，这实际上是一种低信息的先验分布，它能够满足绝大多数实际测验情境。

在修改稿中，我们添加了：“鉴于 $\text{logit}(-1.096) \approx 0.25$，所以该设定与四则一选择题的理论猜测概率相符合”。请见正文第 7 页。

意见 5：

文章 4.2.2 节提到数据的缺失值问题，除假设是完全随机缺失之外，请问是如何处理的？

回应：

感谢您的问题。在修改稿中，我们添加了：“全贝叶斯 MCMC 算法可以根据其他参数的估计值计算出缺失值的后验分布，这是一种“自动填补”的过程，无需做其他设定。”请见正文第 8 页。

审稿人 2 意见：

本文从实际的角度出发提出的一种多阶的认知诊断模型。作者的写作流畅，逻辑清晰，文献详实。文章提出的模型有创新性，有应用价值。但评审人认为，本文仍有一些细节需要修改或补充说明。总体而言，该论文是一篇有意义的模型开发论文，但对模型的应用价值、方式等方面还需做更深入的探讨和说明。

意见 1：

文中的中文、英文的文中引用需统一。英文引用正确，如（Wang & Chen, 2004），中文引用需统一和英文一致，如（刘克文, 李川, 2015）。

回应：

感谢您的意见。根据心理学报参考文献格式的要求，中文作者名之间不添加&符号。

意见 2：

讨论部分，需要作者做出补充，对 MO-DINA 的可能应用范畴做更多的探讨。因为如果一个模型只能满足 PISA 这样一个测验的设计，模型的应用价值将大打折扣。

回应：

感谢您的意见。在修改稿讨论部分中，我们对 MO-DINA 的潜在应用情境做了一些说明和讨论。请见正文讨论部分。

意见 3：

MO-DINA 提出的理论原因是成立的，模型参数设置合理，也确实更适合文中例举的 PISA 数据。但文章在实证研究部分，DIC 和 BIC 并没有很一致的支持 MO-DINA 模型，这当然与实证数据的本身问题和指标问题有关。但评审人仍然期待基于模拟研究的模型比较。

回应：

感谢您的意见。但经过综合考虑，我们认为增加模拟研究部分的比重不适合于《心理学报》现阶段的风格。因此，我们仍把模拟研究视为一个“补充性”的佐证，用于说明 MO-DINA 的参数估计返真性较好。

意见 4：

文章开篇提到现有 PISA 的分析主要使用 IRT 模型完成。尽管 IRT 无法提供评估所需的诊断信息，但对于被试的总体能力信息 θ 仍是可以使用的。而对于 MO-DINA 而言，实际上二阶和三阶都拥有相对整合的 θ。当不关注二阶潜特质对三阶潜特质的贡献问题时。MO-DINA 所估计出来的 θ 值是否比 IRT 估计出来的 θ 值有着更准确地被试排名？类似是否比 HO-DINA 估计的 θ 值更有优势？或是它们的优势在哪？

回应：

感谢您的意见。实际上 de la Torre 和 Douglas (2004)已经对比过 HO-DINA 模型与单维 2PL 模型，发现 HO-DINA 模型中的高阶潜在特质估计值与单维 2PL 模型的能力估计值具有高相关。在修改稿中，我们添加了使用单维 2PL 模型、MO-DINA 模型和 HO-DINA 模型得到的“科学素养”估计值之间的相关关系。发现 MO-DINA 模型与 HO-DINA 模型的相关系数为.998，而 MO-DINA 模型与 2PL 模型的相关系数为.936，表明三者对“科学素养”的估计值具有高相关，三模型所测量可能是同一个潜在特质。请见正文第 10 页。

意见 5：

本文是先做实证研究，后做模拟研究，且模拟研究中并未做模型比较。这可能与找不到合适的第三方数据生成方法有关。建议作者考虑使用一个多步骤的数据生成方式或非参数的生成方法。建议作者仍然将模拟研究提前，实证研究放后。在没有用模拟研究证明模型可行性的情况下分析 PISA 的数据，逻辑上是有问题的。例如文中对 2 号和 23 号被试的探讨，在还未确认属性和能力估计可靠性的基础上，根本无法确认这个差异是否有意义。

回应：

感谢您的意见，我们也认同您的观点。但与上面对您意见 3 的回复类似，未避免增加模拟研究的比重，我们仍仅把模拟研究视为一个“补充性”的佐证，用于说明 MO-DINA 的参数估计返真性较好，同时也佐证了实证研究中的分析结果具有一定的可靠性。

意见 6：

文中有提到，当二阶潜特质显示较高的相关时，不一定就有高阶潜特质，这就涉及到一个模型选择的问题。那么正常应用时除了根据测验编制的理论外，还可以根据那些指标来确定是否选用高阶模型？

回应：

感谢您的意见。修改稿中，我们针对该问题添加了一些简单讨论。请见正文讨论部分。

第二轮

审稿人 1 意见：

对于审稿人提出的几个问题，作者做了详细的、合理的回应。对文章做了相应的修改。审稿人认为这是一个比较好的研究。

回应：感谢您对本文提出的意见和建议。

审稿人 2 意见：

上一轮的修改作者已经较好的回答了审稿人的问题，当前审稿人没有更多的问题了。仅有两个标题方面的修改意见。第一，当前论文的一级标题 4 为“PISA 2015 科学测评数据分析”，建议改为“MO-DINA 模型的实证探索与分析”，因为 PISA2015 作为使用的数据，只是模型的引子，或者例子，作者的意义也并非强调数据的分析接受。第二，当前论文的一级标题 5 为“参数估计返真性探究”，建议改为“MO-DINA 模型参数反正性的模拟研究”。第三，当前论文的二级标题 4.1 为“研究目的与问题”，建议改为“研究问题与目的”，逻辑上而言应该是先有问题，后有目的。

回应：

感谢您对本文提出的意见和建议。修改稿中我们已经按照您的建议对文章做出了修改。

第三轮

责编意见：

意见 1：

题目应当避免多句子结构。目前的题目使用了问号，后面还有句子。建议将题目改为“基于多阶认知诊断模型测评科学素养的一种方法”。

回应：

感谢您的建议，经过思考，修改稿中我们已经将题目修改为“一种基于多阶认知诊断模型测评科学素养的方法”。

意见 2：

除了本文关键词和容易引起翻译歧义的词外，删除正文中的英文概念。例如，教育心理学(educational psychology)，其中的英文概念没有用处。

回应：

感谢您的建议，修改稿中，除一些必要的名词外，我们已经对文中的英文名词做出删减。

意见 3：

摘要中，科学素养为三阶潜在结构，还是科学素养包含三阶潜在结构？请准确措辞。

回应：

感谢您的建议，应是“科学素养包含三阶潜在结构”。我们已经对全文的措辞进行校准。

意见 4：

引言部分可以压缩，删除常识性的、政策性的介绍，按心理学报学术论文的通行做法来行文。

回应：

感谢您的建议，除心理学背景读者外，考虑到一些科学教育和学习科学背景的读者也可能对本文感兴趣，所以原文中的政策性语句更多地是给这类读者呈现的，以便让他们了解 CDA 是符合我国当前教育政策导向的。

修改稿中我们已经对引言部分做了缩减，仅保留了个别政策性文件的语句，为新方法在其他相关学科中的推广和实践应用中的推广提供政策基础。

意见 5：

目前的图 2 没有直观地反映三阶结构。建议参考心理学期刊上有关层次结构的概念图（如自我概念的层次结构）重新画图。但图 6 可以不改。

回应：

感谢您的建议，修改稿中我们已经对图 2 做了修改。

意见 6：

建议删减可有可无的图表，例如，图 5 对本文是必须的吗？

回应：

感谢您的建议，修改稿中已经删除图 5。

意见 7：

数学公式太多，请参考 APA 期刊 Psychological Methods 的文章写法，尽量避免可有可无的公式。不可避免的，除了正文行文需要外，都可以抽出来放到附录中，这样就不会影响读者阅读。这一条不是建议，而是要求，因为心理学报是综合性期刊。不同意修改的话，考虑投稿给诸如 Psychometrika 期刊吧。

回应：

谢谢，您的建议使我们进一步明确当前《心理学报》的学术风格。基于此，修改稿中我们已经对文章结构做出了一定的调整，比如，减少了公式数量、将参数估计等部分移至附录。