

# 《心理学报》审稿意见与作者回应

题目：多维对数正态作答时间模型：对潜在加工速度多维性的探究

作者：詹沛达，满凯文，焦红

---

## 第一轮

审稿人 1 意见：

意见 1：众多研究表明，被试的项目反应时间是很重要的辅助信息资源，所以反应时间的建模与应用是目前心理和教育测量领域的热门研究问题之一。本研究针对现有研究的不足，提出了多维对数正态反应时间模型，并成功应用 MCMC 抽样实现了模型的参数估计。目前关于多维测验下时间模型的开发尚处在起步阶段，该研究具有一定的前沿性，但研究方法上还存在以下不足：1. 模型构建缺乏理论依据和现实意义的支撑。模型应该是实际理论或现象的抽象、概括和描述。本研究提出的多维速度没有一定的理论依据。测验关注的多维能力是多维的，并不意味着被试作答速度也是多维的。论文要阐释清楚为什么被试的速度特质是多维度的，如何对多维速度进行解释，赋予直观意义。不能直接借用多维 IRT 模型的建模策略，提出多维时间模型。多维 IRT 模型是由测验关注能力的多维度而提出的，是有显示依据的，但它不能作为时间模型的依据。所以，要给模型的建构提供切实理由。2. 实际数据分析深度不够。实际数据应重在数据分析本身，基于新的方法有何新的发现，或更加深度地挖掘了数据的信息。模型比较是必要的，此外还要丰富数据分析的角度，提供一些更为有价值的发现。要为模型的实际应用提供有价值的依据和参考。3. 模拟设计需要改进。基于多维模型生成数据，比较单维和多维模型，是不公平的。结果一定倾向生成数据模型。此外，可以判断该模型的估计效果应该没有问题，因为这就是一般的对数线性模型的参数估计问题，参数结构并不复杂。模拟设计可以继续改进。

给出建模的依据，为什么建构新模型是有意义的。因为没有，所以构建，是不充分的理由。被试能力应该是时间长短的影响因素，是否可以在时间模型中加入被试的多维能力呢？如果被试能力显著决定时间，那么说明时间也包含了被试能力的信息，联合使用作答数据和时间数据能够提高对被试能力测量的精度。这样模型的意义就更具体了。此外，实际数据也

可以尝试分析，那个维度的能力对被试答题时间的影响大，那个能力小，再借助 PISA 测试的特征和理论，可以做一些更为深度地分析。

回应：感谢审稿的建议。实际上，原文从理论和数据两个视角为本研究的核心假设(i.e., 建模依据)提供了支撑。首先，原文已经在引言给出了一些认知心理学的证据来支持我们提出潜在速度多维性假设；其次，EFA 结果也表明，RT 数据背后的多维潜在结构与 RA 数据的多维潜在结构是相匹配的。

另外，您所述的“能力应该是时间长短的影响因素”是基于实验心理学的认知，即速度-精度权衡。但实际上，这种权衡反映的是加工速度与加工精度在个体内(within-person)的关系，难以通过一个横断研究/测验来评估。通常，对于一组固定的题目/任务，一旦被试的加工速度固定下来，那么他/她的加工精度也是固定的；因此，建议对加工速度和加工精度分别建模，而与之相对应的潜在加工速度和潜在能力之间的关系可以在更高的层次上建构(van der Linden, 2006; 2007; 2009)。这也是层级建模框架(van der Linden, 2007)的基本逻辑。在修改稿的引言部分，我们已经对此进行了简单说明。

意见 2：建立项目反应变量和反应时间变量的联合模型。Zhan 建立了一个多维测验的包含时间变量的联合模型，他们采用的是 van der linden 的等级建模策略。等级建模能描述被试总体层面的速度与能力的关系，但被试个体层面，两个变量还是独立的。建议作者可以从这个角度提出一个更加一般化的联合模型。

回应：感谢审稿人的建议。我们认为一篇文章关注两个议题是不适宜的，会导致内容冗长，并给读者带来双重困扰。van der Linden (2006; 2007)也是在两篇文章中分别建构了 RT 模型和联合模型。

实际上，您所提及的在一篇文章中同时建构新的 RT 模型和新的联合模型的做法，正是我们的初稿的做法，该文章正在外文期刊外审中。根据两位审稿人的建议，为避免一篇文章含有多个议题导致重点不突出，我们专门把 RT 模型从初稿中摘取出来，撰写了本文。

意见 3：模拟设计如果要验证估计方法的返真性。可以多考虑一些测验条件：例如，不同的题目数/样本量/维度数。还有应验证先验分布的影响。汇报结果可以包含，估计的计算时间和马氏链的收敛效果等。模拟设计还可以从模型的应用价值，例如提高被试的能力精度，模型拟合方法对该模型的选择效果等方面设计模拟。

回应：非常感谢您的建议。鉴于《心理学报》目前对心理计量等方法类文章的定位，依据之前责编的要求，我们借鉴了 *Psychological Methods* 中一些文章的做法(以实证数据分析为主，

模拟研究为辅；这与 *Psychometrika* 等期刊的行文风格所有不同）（e.g., Harring, McNeish, & Hancock, 2016），特意压缩了模拟研究的篇幅。换句话说，本文中的模拟研究是基于实证研究情境进行的，用于辅证实证研究中的结论。该行文风格也被詹沛达，于照辉，李菲茗，王立君 (2019)所使用。

结合您的意见，在不过多增加模拟研究篇幅的前提下，我们在修改稿中增加了一个模拟研究，拟探讨冗余地指定潜在加工速度具有多维性所带来的影响，详见模拟研究 2。

Harring, J. R., McNeish, D., & Hancock, G. R. (2016). Using phantom variables in structural equation modeling to assess model sensitivity to external misspecification. *Psychological Methods*

詹沛达，于照辉，李菲茗，王立君. (2019). 一种基于多阶认知诊断模型测评科学素养的方法. *心理学报*.

.....

#### 审稿人 2 意见：

**意见 1:** PISA 数据公布的数据中所提到的几个 content field 能否认定为 Q 矩阵的元素，尽管作者提供了类似做法的文献，这一做法仍存在争议。如果这个地方的 Q 矩阵并不是 CDA 中的 Q 矩阵，作者需要进一步解释清楚。

**回应:** 感谢审稿人的建议。实际上，Q 矩阵本质就是一个验证性矩阵，用于界定题目与潜在变量之间的关系，广泛使用于多维 IRT 模型、结构方程模型、认知诊断模型等领域之中，其中的潜在变量并非局限于认知诊断中的细颗粒属性。从另一个角度看，CDM 作为一种有约束的潜在类别模型，其中的 class 也并不特指包含细颗粒属性的向量，任何类型的 class 都可以。因此，我们认为不必将 Q 矩阵局限在认知诊断这个小方向上；同时也不必将 CDM 中的“属性”局限在“细颗粒知识、技能”这个小范围里。比如，Li, Kim 和 Xiong (2020)就将潜在变量设定为学习模式(learning mode)，并使用了认知诊断模型进行数据分析。对此，修改中我们也增加了一点说明，详见第 4 页角标 1。

Li, H., Kim, M. K., & Xiong, Y. (2020). Individual learning vs. Interactive learning: A cognitive diagnostic analysis of MOOC students' learning behaviors. *American Journal of Distance Education*.

**意见 2:** 通过探索性因素分析发现的潜在结构能与 Q 矩阵中的结构等同起来？

**回应:** 针对 RT 数据的 EFA，探索的是 RT 数据背后的潜在结构；而 Q 矩阵是用于表达 RA 数据背后的潜在结构的一个验证性矩阵（i.e., Q 仅是一个潜在结构的表达形式）。依据已有认知心理学方面的研究，本研究假设被试的潜在加工速度与潜在能力之间具有相匹配的多维结构。EFA 结果显示，RT 背后的潜在结构与 RA 数据背后的潜在结构（由 Q 所表达）相匹配；即无论从 RT 数据还是 RA 数据看，题目所属的潜在维度是一致的，因此，后续研究我

们直接使用了 Q 矩阵来表达 RA 和 RT 数据背后一致的潜在结构,这也符合我们的理论假设。  
修改稿中我们对此增加了一些解释, 详见第 4 页和第 5 页。

### 意见 3:

下面是一些表达和语法上的建议

- 1、第 1 页第 8 行, 研究者们基于认知/心理学理论, 这种表达不规范, 难于理解
- 2、第 1 页第 21 行, “且”多余
- 3、第 1 页第 22 行, “应”多余
- 4、第 1 页第 23 行, 标注的参考文献处少了右括号
- 5、第 1 页第 26 行, “则”多余
- 6、第 1 页第 28 行, “换句话说, 每个测验维度中的潜在加工速度……潜在能力相匹配的。”表达不清晰。潜在加工速度和潜在能力的主体都是被试, 目前的这种表达容易让人产生误解。
- 7、第 1 页第 29 行, 举例应和前面的表述一致, 直接表达成“被试的某方面的潜在能力是与这方面的加工速度相匹配”较好。
- 8、第 2 页 12 行和 13 行, 对 dimension-specific processing time 的两个不同的翻译是有意为之? 感觉不到这么做的意义, 或者是出现了 typo
- 9、第 5 页 6 行 MRM-MLRTM——MRM 指什么, 上下文没有找到对应解释。MRM-MLRTM 也就给出明确的定义。
- 10、第 5 页第 25 行, ppp 也需要列出英文全称。对 PPMC 而言选取一个合适的差异测度“的” (需修改) 必要的
- 11、第 6 页第 12 行, 题目时间强度参数“个” (多余) 的估计
- 12、第 6 页第 19 行, 句末少了“。”
- 13、第 8 页第 6 行, 马尔科夫链数“,”改“、”
- 14、第 8 页 16 行 需要明确 bias 和 RMSE 这里说的是强度参数还是区分度参数。 1
- 15、作者在文中 bias 写法不统一, 有的是 bias,有的是 Bias
- 16、第 9 页表 7 的表标题有问题
- 17、第 9 页第 15 行, 句末少了“。”
- 18、第 11 页 27 行, 引用的文献位置不正确

回应: 感谢审稿人的认真审阅。修改稿中我们除了对上述内容进行了修改, 还认真通读了全文, 尽量避免出现类似小错误。

---

## 第二轮

### 审稿人 1 意见：

作者较好的回答了审稿人的问题，也做了相应的补充，很好的解释了多维反应时间模型的应用价值，此外该研究丰富了反应时数据建模方法体系，具有一定理论价值。该研究还存在以下问题，建议修改：

1. 模拟研究和实际数据分析可调整先后顺序。
2. 模拟研究 1 中关于参数返真性比较，数据生成条件中被试多维潜速度之间的相关性只设定了一种情况，相关系数为 0.15，是否可考虑多个相关程度，初步预判，随着潜在速度相关性的增强，单维时间模型和多维时间模型参数的返真性会越来越接近，但这不影响本文研究价值，会增强模拟研究的客观性和全面性。毕竟，弱相关下，多维模型好是正常的。此外，PISA 数据分析中，潜在速度的相关度很高，在 0.7 以上。所以，模拟中多考虑一些相关程度与实际数据分析还可前后呼应。
3. 实际数据分析，PISA 的维度定义，属于题目间多维，而不是题目内多维，在本文的方法介绍部分，提到了两种多维测验的定义，作者在实际数据分析部分可简单对此进行补充性解释。此外，所选 PISA 测试的题目数是 9 道，这么少的题目数是否会影响分析的准确性和可靠度？可简单说明，使分析更客观一些。

回应：感谢您的肯定与建议。

首先，为了保证 MLRTM 可以使用基于 RA 数据建构的 Q 矩阵，我们需要先明确 RT 数据的潜在结构是否与 RA 数据的潜在结构(i.e., Q 矩阵)相匹配。因此，先进行实证数据分析可能更符合研究逻辑。

其次，在模拟研究一种，多维潜在速度之间的相关系数被设定为 0.6（而非 0.15），是中等程度相关。该设定参考了实证数据分析结果，具有实践意义。我们未过多地增加模拟研究内容的主要原因在上轮审稿中已经给予说明(i.e., 我们希望增加方法类研究中实践数据分析的比重)。当然，如您所述，增加自变量条件确实会让模拟研究结果更为丰富，这点我们在讨论部分做了进一步说明。

最后，感谢您的建议，在修改稿的讨论中我们补充了相应的内容。

### 审稿人 2 意见：

本研究作者改进了已有的模型，构建了一个基于反应时的题目内多维的模型。受计算机技术的发展和普及，测验越来越趋向于计算机化，反应时数据的收集变得容易和自然。因而催生了很多基于反应时的研究和应用。反应时数据的信息它不但反映了被试的能力，同时也反映了被试的认知加工过程，因此，纳入反应时数据的信息对于心理和教育测量是非常有意义的工作。建议录用。

回应：感谢您对本文的肯定。