《心理学报》审稿意见与作者回应

题目：计算机动态测验中问题解决过程策略的分析：多水平混合 IRT 模型的拓展与应用
作者：李美娟　刘玥　刘红云

第一轮

审稿人 1 意见：

意见 1：本文试图对教育测量中的过程性数据(Processing data)进行分析，这类研究在国内还不多。但是，从国外来看，对过程性数据的分析并不新鲜，作者需要对相关的研究进行介绍。文中存在一些表达不够准确的地方，已经在文标出，请作者进行仔细修改。

回应：谢谢审稿老师的建议！我们已经在每一处批注处进行了详细修改。请详见修改稿。具体修改内容如下：

① 第 1 页引言第 1 段第 5-8 行进行了相应举例。

② 第 2 页引言第 2 段第 5-6 行填加了英文名称和参考文献。以及第 3 页引言最后一段第 2 行填加了参考文献。

③ 第 2 页引言第 4 段对传统的 MMixIRT 模型进行详细的介绍，包括模型的数学形式。

④ 第 4-6 页 2.1 模型定义部分对模型进行了详细解释。

⑤ 请作者解释清楚在目前的模型下，过程水平和个体水平模型之间关联是什么？是如何产生联系的。

在公式（2）和（3）中当 $\boldsymbol{y_{ki} = y_{Jki}}$时，如果同时估计个体水平和过程水平模型,学生最后一步作答数据既用于过程水平模型的估计，同时也用于个体水平的模型估计。因此，C_{Jk}，和 θ_k 是存在关联的。之前在文中讲得确实不是很清晰，我们已经在修改稿中进行了明确标记。详见第6页第2段。

⑥最后一步的作答数据是一个 0 或者 1？作者的意思只用这一个作答来估计被试的能力吗？

我们在修改稿第 10 页第 1 段中加入了表 4，呈现了过程性数据的编码方式，举例说明第一层变量的放置方法。学生最后一步的作答数据是在 23 条路径上的最终作答状态，是 0 或 1 的状态。拓展的 MMixIRT 模型使用学生最后一步的作答数据来估计被试个体层面的能力值。与目前传统的被试在一道题目上的做答只记为对错相比，基于目前模型到的个体层面的能力估计可以为错误做答被试提供更多的信息。

⑦什么是 SPMF 平台？

SPMF平台是一个采用Java开发的开源数据挖掘平台。网址 http://www.philippe-fournier-viger.com/spmf/index.php?link=documentation.php。我们在修改稿的相应位置进行脚注标明。

⑧“将每个学生最后一步所选路径属于的策略作为其最终的策略，然后统计各策略的能力估计值平均值”这句话让人很费解，最后一步都是要到达 Elnstein，从图上看，可以从 Mandela 到 Elnstein,也可以从一个没有名的地方(Nowhere)到 Elnstein。如果是这样，何来 5 种策略。否则这句话需要修改正确。

之前这里的解释比较简单，不容易理解，修改稿已经在第 13 页第 1 段进行了详细说明。

意见 2：从作者所介绍的模型来看，作者采用了潜在类别分析对被试进行了分类（即不同作答策略），采用 IRT 对被试进行了潜在特质分析（即不同潜在特质）。按照这个思路，本研究

中的数据也可以采用高阶认知诊断模型进行分析，同样可以得到作者类似的结果和结论。请作者解释，这二者的差异是什么？

回应：谢谢审稿老师的提问！您的提问使得我们对模型的理解更加深刻！在查阅文献的基础上，我们对两个模型进行了总结和对比，

不同点：高阶认知诊断模型的本质是一种可以考虑测验层级结构的认知诊断模型，因此属性标定是模型分析的基础。高阶认知诊断模型的高阶主要指的是测试结构维度之间的阶层关系，而不是由于数据不满足独立性而带来的嵌套关系；高阶认知诊断模型最低一层的分类主要的针对属性掌握的分类，是基于被试在测试不同属性的题目的表现估计属性掌握程度的分类，高阶是不同层的能力的估计。关于高阶认知诊断模型，近年来 de la Torre 和 Douglas（2004），詹沛达、于照辉、李菲茗、王立君(2019)等近年来作了比较深入系统的研究，詹沛达等还将其用于科学素养的测评（详见心理学报 2019 第 5 期），更为详细的内容可以参考具体的研究。多水平混合 IRT 模型的本质是考虑数据之间嵌套结构的多水平分析模型，解决不同次操作数据之间不独立的问题是其出发点。即多水平主要是过程嵌套于个体的数据层级结构，类似于追踪研究中的重复测量嵌套于个体；过程数据的分类主要是基于被试在完成一个任务过程中不同操作特征体现出的类别特征，是一种数据驱动的分类思想，不需要事先做属性特征标定，但分类后是需要根据类型所表现的特征解释类别的特点，即本研究中给出的其在完成一个任务过程中选用策略的不同的表现。

相同点：两个模型对低阶（或低水平）的分析都是类别的划分和类别特征的描述，高阶（或高水平）都是估计个体的能力特征。对于目前研究的数据，如果采用认知诊断模型，需要有两个前提，一是基于过程数据能否界定明确的属性，并进行标定，结合目前本文分析的任务，由于任务本身比较简单，可能并不需要做多个属性标定；二是需要再认知诊断模型中考虑数据的嵌套结构，多水平认知诊断模型（Wang & Qiu, 2019）可以提供一些借鉴。也可以作为未来研究的一个方向。

参考文献：

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

Wang, W. C. , & Qiu, X. L. . (2019). Multilevel modeling of cognitive diagnostic assessment: the multilevel dina example. Applied Psychological Measurement, 43(1): 34-50.

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, 51(6)*,734-346.

我们在修改稿中对认知诊断模型进行了总结和分析，请详见引言部分第 3 段。具体修改内容如下：

近年来，关于问题解决过程以及解决问题过程中所需技能和策略的探讨，随着测量理论和统计技术的发展越来越被重视。其中一类是基于测验题目信息，通过对完成该题目过程所需的技能（或属性）进行标定，基于一定的测量模型对解决问题过程的策略特点进行分析。最具代表性的方法是基于认知诊断模型的评估。如 de la Torre 和 Douglas（2004）采用高阶潜在结构模型，对学生能力进行估计，并基于学生的认知属性掌握模式对学生的认知特征进行分类。詹沛达、于照辉、李菲茗、王立君（2019）在其基础上提出多阶认知诊断模型，并将其用于科学素养的测评。通过对题目进行认知技能（属性）标定，基于题目的属性矩阵（Q 阵）和学生的作答反应对学生的属性掌握模式（即实现对学生的分类）和能力进行评估。另一类则是基于信息技术获得的丰富的过程数据，借助于统计模型和数据挖掘的思想，对过程数据蕴含的丰富信息进行分析。常用的方法有可视化分析方法（DiCerbo, Liu, Rutstein, Choi, & Behrens, 2011）、聚类分析方法（Bergner，Shu, & von Davier, 2014）和分类分析方法

（Desmarais & Baker, 2012)。最近，也有学者（Shu, Bergner, Zhu, Hao, & von Davier, 2017）结合隐马尔科夫链模型（Hidden Markov Model）和项目反应模型，来分析过程性数据中的序列作答信息，从而估计学生的能力。本研究探讨的方法和研究的重点属于第二类，即基于过程数据分析解决问题过程中学生所用的不同策略类别，同时基于任务提交状态的信息对学生的能力进行估计。

意见 3：作为一个拓展的模型，作者需要先对模型的参数返真性进行探讨，需要介绍参数估计所使用的相关算法。如果各模型参数都能够很好的估计，然后才是对实际数据的分析。
回应：谢谢审稿老师的建议！我们在修改稿第 6-7 页增加了 2.2 参数估计的返真性与分类准确性部分。

意见 4：请作者提供相应的 Mplus 代码。
回应：谢谢审稿老师的建议！我们已经在修改稿附录 1 中填加了 Mplus 代码。

审稿人 2 意见：
意见 1：文章采用多水平混合 IRT 模型对计算机动态测验中问题解决过程策略进行分析，研究问题较为前沿，方法选择得当。针对动态过程性数据的分析，目前研究尚处于起步阶段。文章所采用的研究方法也能恰当地解决过程性数据的问题，具有开拓性价值。建议文章做出一下修改和调整，以便更好指导其他研究者。就整体研究问题而言，研究者所关注的重点为“问题解决过程”。本文亦属于应用型研究而非单纯方法类研究。建议作者对文章结构稍作调整，在引言部分对“计算机动态测验”“问题解决的过程性数据”“策略选择”等领域的文献做拓展和补充。
回应：谢谢审稿老师的建议！ 我们已经对引言部分进行了修改，对“计算机动态测验”“问题解决的过程性数据”“策略选择”进行了文献的拓展和补充，请见引言的第 1 段和第 2 段。

意见 2：第 5 页，第 1 段。本段引入了多水平混合 IRT 模型。但是前文并没有对该模型做简要介绍。特别是混合模型的使用，作者应阐述清晰。
回应：谢谢审稿老师的建议！修改稿中引言第 4 段话中增加了对“传统多水平混合 IRT 模型的介绍以及使用”。

意见 3：第 5 页。图 1 和表 1，地图有误。如，没有“nowhere”地名的标记（左侧第二个地点的位置？）。建议作者勿直接截图，应重新作图。
回应：谢谢审稿老师的建议！这个地方我们之前确实没说清楚。因此，我们对图 1 的图注进行了修改：地图上两个节点之间的路线为一条路径，标蓝的路径为正确路径。地图中没有标记地名的地点均称为 nowhere。

意见 4：第 6 页。对过程性数据编码的方法，作者没有阐明清晰，特别是针对表 2 的说明。首先，反应时在小数点第二位开始已经全为 0，所以没有必要保留小数点后四位。其次，路径选择一列和后面 P1 至 P23 的数据标记未对应，比如第 1 行，路径选择一列为“01000……”则应表示选择 P2，但是后面以及说明过程均显示选择 P1。
回应：谢谢审稿老师的建议！首先，我们已经将表 2 和附表 1 反应时和其他变量中小数点后面多余的 0 全部删掉。另外，表 2 中路径是我们写错了，已经进行修正，“例如，第一行表示编号为 00017 的学生在第一步选择 P2，第二步选择 P1，第三步选择 P13，……，第 8 步

取消 P1，第 9 步取消 P18，……”。具体修改内容详见第 9 页 3.2 过程性数据编码部分。

意见 5：第 7 页，第 1 段。请详细说明表 2 的数据是如何编码至模型数据的。文章没有对模型的原始数据进行详细的说明，特别是第一层数据的格式。此外，就多水平 IRT 模型而言，第一层变量的放置方法没有详细说明，而这恰恰是本文的创新之处，建议举例并详细说明。
回应：谢谢审稿老师的建议！我们已经在文中第 12 页第 1 段详细举例说明第一层变量的放置方法，并进行举例，增加表 4，即表 3 的数据进行编码后的结果呈现。

意见 6：文章缺少“研究方法”中变量选择的介绍部分，更像是一个探索性的分析。如第 16 页的“路径点击数”等。这样降低了阅读连贯性。建议作者将后文中所有涉及到新的概念出现的问题，专门在“3 研究方法”后面加一个“3.3 变量”的说明，避免在后期再出现新概念。
回应：谢谢审稿老师的建议！我们已经在修改稿中第 11 页第 2 段中增加了 4.2 变量说明部分。

意见 7：第 10 页。最后一段。五个地区的数据是分开估计的，还是放在一个模型下估计的？
回应：谢谢审稿老师的提问，模型对五个地区的数据是同时估计的。我们在原文中确实没有说明，修改稿第 12 页第 1 段第 2 行中，增加了相应的说明。

意见 8：第 15 页。最后一段。行文中“表 8”应为“表 9”？
回应：谢谢审稿老师的建议！我们确实写错了，已经在修改稿中进行了修正。

意见 9：第 17 页。建议将表 9 进行优化。比如小数点位数的保留是否有效，有大量“0.000”的单元格可否考虑优化等问题。
回应：谢谢审稿老师的建议！我们已经将表 9 以及所有表格中出现类似情况的数据进行优化。

意见 10：第 15~16 页。对路径点击数等概念的说明，应放在方法部分。详见 6。
回应：谢谢审稿老师的建议！这部分内容我们已经修改到方法部分，请参见意见 6 修改内容。

意见 11：文章在行文过程中有用到“31 分”这个概念多次。建议作者将这个概念改为“正确作答”。因为文章包含题目中的时间、学生反应时、耗时等类似的概念，不容易区分，也加深了阅读难度。“耗时”是指学生作答反应时还是题目中所耗用的路径时间？请在方法部分将各个概念标识清晰。
回应：谢谢审稿老师的建议！原文中对于题目中的时间、学生反应时、耗时的概念描述确实不清晰，我们已经根据您的建议将“31 分”修改为“正确作答时间”。且在 4.2 变量中进行了重新描述和阐述。参见意见 6。

意见 12：讨论部分目前的论述尚可。但希望作者加入不同国家之间异同的讨论。因为有 5 个国家和地区，之间有较大的差异。可以引用适当的文献，如李瑾教授对中西文化差异的研究、OECD 中中西方文化学生的差异，包括他们的学习策略、学习动机的异同等。
回应：谢谢审稿老师的建议！我们已经在讨论第三段中加入了不同国家之间异同的讨论。详见讨论第 20 页第 1 段。

第二轮

审稿人 1 意见：

意见 1：作者对原模型进行拓展的意义是本模型比原模型更能适应某些情形，所以，研究中需要提供的重要内容应该是对两个模型的分析结果进行对比，或者说明只有本模型才能适应当前情形，其他模型无法在此情形下运行，以说服读者。

回应：感谢审稿老师的建议！原文这一部分的叙述的确不够清楚。我们在修改稿中更加清晰地描述了传统的 MMixIRT 模型，同时补充了其在处理过程数据时的不适用之处。除此之外，在介绍拓展的 MMixIRT 模型时，我们强调了其与传统模型的不同之处，将其作为单独的一小节进行了新表述。请详见修改稿第 3 页最后一段和第 4 页第一段，以及第 4 页和第 5 页的第二部分。

本研究的主要目的是将传统的 MMixIRT 模型进行表述上的修改和拓展，以能够适用于分析问题解决测试获得的过程数据。主要修改内容和更为详细的解释如下，供您审阅。

2. 拓展的 MMixIRT 模型

结合过程性数据的嵌套结构，对传统 MMixIRT 模型做了两方面的修改和拓展。

首先，步骤的累计信息作为特定步骤的过程数据更能体现问题解决任务过程数据的特点。传统的 MMixIRT 模型假设每个时间点的测量都是某一特质在这一状态的反应，过程数据中某一时间点的反应只是这一时刻所执行的一个行为操作，如果只采用这一步操作的信息，无法体现行为序列的连续性。这一步骤可以在数据整理阶段进行实现。模型可以表示为：

$$Y_{jki} = \sum_{t=1}^{j} w_t \, y_{tki} \qquad (2)$$

其中y_{tki}为第 k 个学生 t 时间点在 i 路径上的操作行为。传统的 MMixIRT 是直接对y_{tki}建模，而拓展的 MMixIRT 是对累计反应Y_{jki}进行建模。如果对于时间 t=j，$w_t = 1$，否则$w_t = 0$，则变为传统 MMixIRT 模型。结合所采用的测试题目和过程数据的特点，采用累积反应做答作为过程 j 的反应作答，即如果$t \le j$，则$w_t = 1$。

其次，过程层面和个体层面变异的分解增加设计矩阵 A，可以使得模型对过程水平和个体水平变异的分解更加灵活。在传统的 MMixIRT 模型中，$Y_{jki} = Y_{jki}^{(w)} + Y_{ki}^{(B)}$，即将$Y_{jki}$的变异分解为第一水平（组内$Y_{jki}^{(w)}$）和第二水平（组间$Y_{ki}^{(B)}$）两部分。拓展模型相当于在原有模型的基础上增加了一个设计矩阵A，每个被试对应设计矩阵的行数为记录的过程行为的次数，列数为 2。模型可以表示为 $Y_{jki} = A_j \begin{pmatrix} Y_{jki}^{(w)} \\ Y_{ki}^{(B)} \end{pmatrix}$ （3）

其中 A_j 为设计矩阵的第 j 行，用来定义过程数据不同层面潜变量的分解权重。如果对任意的 j，设计矩阵$A_j = (1,1)$，则是传统的 MMixIRT。

可以看出，传统模型是拓展模型的特例。拓展模型和传统模型的区别主要表现在以下两个方面：（1）过程水平每一步骤的潜在类别是前面各个步骤累积的状态表现，而不是这一个步骤的表现，描述累积状态不仅可以更好地解释解题过程策略的使用，而且可以为探索策略使用的连续性和策略的转换提供依据；（2）个体水平潜变量的定义所采用的测量指标与传统的 MMixIRT 模型不同。传统的 MMixIRT 模型中，个体水平的潜变量是由第一水平的观测变量[$y_{jk1}, \ldots, y_{jkb}, \ldots, y_{jkI}$]估计得到（Lee, Cho, & Sterba, 2017），而拓展模型中可以定义更加自由的设计矩阵 A 决定个体层面能力估计所用到的信息。

3.本研究使用的拓展 MMixIRT 模型

拓展的 MMixIRT 模型比较灵活，可以在第一水平和第二水平模型中结合实际研究关注

的重点定义不同的模型。结合过程数据的特点，本研究主要关注学生在问题解决过程解题策略的差异和最终状态体现出学生个体能力的差异，因此，本研究所使用的模型也是上述拓展模型的特例。

关于模型的优劣，本篇文章的核心不在于比较传统模型和拓展模型哪个更优。主要是因为模型的选取首先应该根据数据的特点，选择适合数据特点、能够回答所关心问题的模型。我们没有比较两个模型的结果，因为我们认为两个模型描述的数据特征不同，潜变量的意义也就不同，其结果也不具有可比性。本研究的主要目的是建构一个适合于描述过程数据的模型，从模型应用的结果来看，模型很好结合了过程性数据的特点，得到了合理的可解释性的结果，例如结果 5 验证了本研究模型建构的合理性。

意见 2：根据作者提供的关于本模型和已有模型之间的区别，我无法判断哪个模型更优。根据第一点区别，我倒觉得已有模型从逻辑上更优，因为他更能体现诊断的意义；根据第二点，我也认为已有模型更优，因为他能利用更细致的数据信息。当然，我的这两点认识可能是错误的，若如此，则请作者从逻辑上解释清楚这两点的具体优势是什么，而不是只说他们不同。

回应：谢谢审稿老师的建议！我们在修改稿中重新对原模型、拓展模型和本研究使用的模型进行了梳理。对于原模型，我们更加清晰地对模型进行表述，同时提出了其在处理过程数据时的不适用之处。详见修改稿第 3 页最后一段。对于拓展模型，我们从模型表述、和传统模型的对比来进行详细说明，并提出传统模型是拓展模型的特例。详见修改稿第二部分（第 4 页和第 5 页）。由于我们提出的拓展 MMixIRT 模型比较灵活，可以在第一水平和第二水平模型中结合实际研究关注的重点定义不同的模型。结合过程数据的特点，本研究主要关注学生在问题解决过程解题策略的差异和最终状态体现出学生个体能力的差异，因此，我们提出了本研究使用的模型，本研究所使用的模型也是上述拓展模型的特例。详见修改稿第三部分（第 5-7 页）。

在处理过程数据中我们所定义的模型与传统模型存在区别，之所以对传统模型进行拓展，最主要的原因是想要结合数据特征建构更合理和更具有可解释性的模型，而非比较两个模型的优劣。第一点是结合过程数据本身的特点以及操作过程之间行为连续性特点的处理，如果只用这一个时间点的操作行为，那么在本研究的数据中，每一次只有一条路径上为 1 或者-1，最终的分类必然是操作相同的分为同一分类（比如分为 23 个类，每一类表示选择了一条不同的路径），这实际上不是我们想要的策略的分类。本研究的重点是关注问题解决过程中的策略类型、策略使用特点以及策略转换，这一处理方式没法达到这一目的。

第二条涉及到我们对个体能力的定义，在过程数据中，用最后提交的状态（或者前面所有行为的累积状态）作为个体能力的估计是合理的，而不需要采用每一步操作信息。实际上采用每一步的操作信息估计得到的潜在能力是当前操作状态的能力估计，反映的是过程步骤的能力，本研究中过程的策略分类一定程度上包含了这些信息，可以从表 7 的结果看出，过程策略与个体能力之间的关系。

总之，这部分模型的选择和定义主要是结合过程数据的特点，选择能够描述数据特点的模型，而非比较两类模型。

意见 3：可能是作者疏忽，作者在介绍模型时，将路径数固定为 23；另外，如果根据作者的解释，个体水平模型中的也许不需要 i 这个下标，根据我的理解，同一个问题的最终目标是同一的。我不知道这个理解是否准确。

回应：谢谢审稿老师的建议！确实是我们的疏忽，在介绍模型时，将路径数固定为 23。我们已经在相应地方进行了修改。另外，因为在个体模型中，y_{ki}表示第 k 个学生在第 i 条路径上的作答。α_i表示第 i 条路径的区分度参数，β_i表示第 i 条路径的难度参数(i=1,2，…，I)。

在模型的表述中，i 表示第 i 条路径，这个下标是必需的。

审稿人 2 意见：

意见 1：本文经过一轮修改，对审稿人的意见有详尽的回复，质量有了明显提高。对多水平混合 IRT 模型的介绍、实证研究的介绍更为详尽。但经过一轮修改，增加了不少内容，建议再对文章的连贯性进行修订。文章的一级标题需要重新考虑。现在的文章，包含了一个模拟研究和一个实证研究。模拟研究在“2”下面，但实证研究包含了“3”“4”“5”三个标题，显得两部分不平衡。可将内容较少的“3”和“4”整合（相应的标题名称，可再做斟酌，作者参考和定夺）。

回应：谢谢审稿老师的建议！我们已经对“3”和“4”的内容进行了整合，详见修改稿。

意见 2： 细节问题，一些缩写、简称行文不符合 APA 规范。比如 2 页，2 段，倒数第 4 行，“VOTAT 策略”。文中再无提及到该策略，中文名称即可；或列出相应的全称。类似的问题包括同页的“OECD（世界经合组织，全称）”；10 页，3.3 样本介绍中，反应时为 669.22s（秒）；11 页“SPMF 平台（全称？）”等。不一一赘述，请作者认真排查。

回应： 谢谢审稿老师的建议！我们已经对“VOTAT 策略”、“OECD（世界经合组织）”反应时为 669.22s（秒）”等细节问题进行修改，并对全文内容进行了认真排查和修改，详见修改稿。但是需要说明的是 11 页 SPMF 平台中的“SPMF”就是平台的名称，不存在中文全称。

意见 3： 附录提供的 Mplus 语句，可在适当部分加上注释（用“！”语句写明亦可）。

回应： 谢谢审稿老师的建议！我们已经在语句中加上注释，详见修改稿。

第三轮

编委意见：

意见 1：目前的论文太长，建议作者将不重要的内容删减。对于示例数据（表 3、表 4），如果不影响理解，可以用少数几行示例便可。

回应：感谢编委老师的建议！我们在修改稿中不仅对表 3 和表 4 进行了删减，同时也对其他图和表以及文章中不重要的内容进行删减。

意见 2：第 4 节标题“研究方法”引入实例，似乎更像案例？如果是研究方法，是什么问题的研究方法？建议检视整篇文章的结构并修订标题。

回应：谢谢编委老师的建议！我们在修改稿中对文章进行精简和调整，同时也对标题进行了修订。