《心理学报》审稿意见与作者回应

题目：问题解决任务中行动序列的二分类建模：单/两参数行动序列模型
作者：付颜斌，陈琦鹏，詹沛达

第一轮

审稿人 1 意见：

意见 1：本研究将问题解决任务中的 2-gram 行动序列进行正误划分，类比于传统测验中的“题目”，并提出采用 Rasch 和 2PL 模型对过程“题目”建模，以估计被试的潜在能力水平。在利用新型心理测评手段时，如何利用过程数据评估被试的潜在能力是心理测量领域的前沿问题，也是一个难题。这个难题不仅仅出现在研究领域，在实践应用领域也很突出。作者非常清楚的呈现了研究问题，将研究关键点集中在如何利用二分类 Logistic 模型对过程数据建模，因此选题比较有针对性。该研究在模拟研究阶段，比较了各种模型在不同样本量和序列长度下的表现，可以为后续的实践工作带来更多的借鉴。在阅读该研究中，还有几个小问题，期望和作者进行深入探讨。作者在引言部分提到，“特征提取可分为理论驱动和数据驱动两种方法”，本研究采用的特征提取方法属于哪一类？如何论证所提取的特征与潜在能力之间的联系？是否具有理论支撑？

回应：

感谢您的问题。修改稿中已指出“与 SRM 一致，ASM 也属于结合随机过程思想的心理计量建模方法。”见第 8 页 3.2 节。

意见 2：引言部分，作者用较多篇幅介绍了 Han 等人(2022)的多分类建模方法，并认为“对于有正误之分的数据，二分类建模更为适宜”，从而提出本研究的研究目的，即对包含正误信息的行动序列进行二分类建模。但以往研究中不乏对于行为序列的二分类建模探讨，建议作者更详细地解释本研究与其他研究的异同之处。部分研究供作者参考：

Fu, Y., Zhan, P., Chen, Q., & Jiao, H. Joint modeling of action sequences and action times in problem-solving tasks. PsyArXiv. Retrieved from psyarxiv.com/e3nbc

Xiao, Y., & Liu, H. (2023). A state response measurement model for problem-solving process data. Behavior Research Methods, online.

Han, Y., & Wilson, M. (2022). Analyzing Student Response Processes to Evaluate Success on a Technology-Based Problem-Solving Task. Applied Measurement in Education, 35(1), 33-45.

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. Psychological Test and Assessment Modeling, 59(1), 109.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. Psychometrika, 85(4), 1052–1075.

回应：

感谢您的建议。我们在修改稿的引言和 3.2 节“与相关模型的对比”中添加了相关内容。比如，在引言中我们补充“另外，还有研究提出了结合随机过程思想的心理计量建模方法，例如，马尔可夫 IRT 模型(Shu et al., 2017)，马尔可夫决策过程测量模型(Lamar, 2018)，连续时间动态选择模型(Chen, 2020)，序列作答模型(Han et al., 2022)以及状态作答测量模型(Xiao

et al., 2023)。通常，这类方法假设在给定潜在问题解决能力的前提下，被试的行动序列之间满足条件独立性假设；比如，将行动序列看作具有一阶马尔可夫特性的离散随机过程(Han et al., 2022; Xiao et al., 2023)，从而在保留序列本身顺序信息的同时推断出连续的潜在能力估计值。然而这些模型仍然存在一定局限性，比如，马尔可夫 IRT 模型以具体操作之间的转移作为“题目”，依各“题目”出现的频次，把具体操作序列转化为频数矩阵进行分析；在此过程中行动序列的顺序信息并没有得到完整保留。马尔可夫决策过程测量模型针对不同的任务情境需要单独为不同的操作定义奖励函数，应用成本较高。连续时间动态选择模型仅考虑了任务整体难度对状态转移概率的影响，难以深入行动层面探究任务的特征。后续，Xiao 和 Liu(2023)在连续时间动态选择模型的基础上提出了状态作答模型，该模型考虑了任务中不同状态转移的独特性，假设不同状态转移具有不同的难度(容易度)参数；但该模型也假设同一状态下所有错误状态转移的发生概率相等且均分于错误状态转移的数量。”

在 3.2 节中，我们补充“其次，除 SRM 外，Xiao 和 Liu(2023)提出的状态作答模型也采用了多分类建模。表 1 呈现了 SRM、状态作答模型和 ASM 之间的对比。首先，鉴于 SRM 和状态作答模型均为多分类建模，两者均涉及各“选项”的发生概率，差异在于 SRM 允许各“选项”的发生概率存在差异，而状态作答模型假设它们相等且均分于错误“选项”的数量；因此，状态作答模型可视为 SRM 的约束模型。其次，当任务中所有问题状态的可选项数量均为 K = 2 时，三个模型完全等价。另外，同样值得注意的是，由于状态作答模型与 SRM 类似也对部分模型参数进行了约束，导致其待估计参数的数量并不总是多于 ASM。

此外，还有个别过程数据分析研究也使用了与 1P- / 2P-ASM 类似的单参数或两参数 IRT 模型的形式。比如，Han 和 Wilson (2022)将混合 Rasch 模型或混合分部评分模型应用于过程数据分析，不仅能够估计学生的潜在能力，还能够对学生的问题解决过程进行探索性分类。Shu 等人(2017)提出的马尔可夫 IRT 模型同样具有与 2P-ASM 类似的两参数 IRT 模型(或分部评分模型)形式。但上述两个模型与 ASM(以及 SRM 和状态作答模型)有两个主要区别：(1)上述两模型的分析对象是具体操作序列，而 ASM 的分析对象是问题状态转移序列，具体操作序列和问题状态转移序列在实际应用中可能存在差异。比如，本文的实证研究例子中(图 6)，两个不同的具体操作序列“城市地铁” →“全价票”和“城市地铁” →“优惠票”对应了相同的状态转移序列“错误的交通类型”F→“错误的折扣类型”G；(2)上述两模型分析的数据是由行动序列转化得到的具有标准化数据格式的数值型矩阵，而 ASM 分析的数据是保留了时序信息的且有个体间长度差异的非标准化格式数据。比如，前者为保证所有被试具有相同长度的数据，常把重复出现但具有前后时序的相同具体操作序列转换为频次信息并使用多级评分模型进行数据分析，但该转换损失了过程数据中重要时序信息。

表 1 三种行动序列数据分析模型的对比

模型	正确行动序列	错误行动序列			
	1	2	3	…	K
序列作答模型	P_1	P_2	P_3	…	P_K
状态作答模型	P_1	$(1-P_1)/(K-1)$	$(1-P_1)/(K-1)$		$(1-P_1)/(K-1)$
行动序列模型	P_1	$1-P_1$			

注: 当前问题状态共包含 K 个可选项(即可形成 K 个行动序列)，其中第一个可选项为正确行动序列，其余可选项为错误行动序列; P 为发生概率.

意见 3：1P-ASM 和 2P-ASM 本质上是 Rasch 和 2PL 模型，请问在应用 1P-ASM 和 2P-ASM 时，是否要对 IRT 模型的前提假设进行验证？比如单维性、局部独立性等？特别的，如何确

保反应序列中的状态转移之间是相互独立的？学生在测验后期的行动是否会受到前期行动的影响？若多次重复同一行动，其正确作答概率不会改变吗？

回应：

感谢您的问题。本研究遵循 SRM 和状态作答模型(Xiao et al., 2023)的假设，在给定被试潜在能力后各相邻阶段呈现的行动序列之间满足条件独立；且在修改稿中已经补充相应的说明，见公式 5 上方。另外，在研究局限中我们已经指出“然而，在一些问题解决任务中，有可能需要被试使用多个不同维度的问题解决能力。后续研究也可尝试进一步提出多维行动序列模型(Shu et al., 2017; 韩雨婷, 2021)。”

此外，本研究提供了绝对模型拟合优度指标后验预测 p 值(ppp)的计算结果，ppp 值接近 0.5 表示模型拟合数据，拟合优度检验的结果能够为模型的局部独立性假设和单维假设提供一定的支持。

意见 4：最优序列是如何定义的？当任务中有多条最优路径时，要如何处理？会不会有些操作在当下是最优的，但是从整体来看并非最优？

回应：

感谢您的提问。修改稿中，我们已经补充“把达到任务目标的最短问题状态转移序列界定为最优问题状态转移序列”，见 2.1 节。

此外本研究提出的模型适用于结构良好(well-defined)的以有限状态自动机(Finite-state automata; FSA)为原型构建的问题解决任务，在这类任务当中，问题解决方案是明确的，满足整体最优的最短路径被定义为最优序列，偏离最优序列或不满足整体最优的状态转移则被定义为错误的状态转移，因此，在相邻序列的定义过程中不存在局部最优而整体非最优的情况。

意见 5：关于模型的应用范围，仅适用于问题解决任务吗？能否应用于其它心理测验的过程数据分析？一些更具体的问题请见审改稿。

回应：

感谢您的问题。修改稿中，我们已补充“本研究聚焦于任务目标明确且已知信息完备的结构良好(well-defined)任务；这类任务常以有限状态自动机(finite state automata)为原型构建。这类任务通常拥有有限的问题状态，有限的用户输入信号(即行动或操作)，并且通过用户的操作可以产生对应的输出信号，即拥有明确的状态转移规则(Buchner & Funke, 1993)。”，见 2.1 节。因此，如果其它心理测验符合 FSA 任务的基本特征，产生的过程数据拥有同样的数据特征，那么本文提出的模型同样适用。

意见 6：审阅稿中其他细节问题。

回应：

感谢您的细心审阅，针对深越稿中的批注，我们均逐一进行了调整和修改，详见修改稿。部分未修改的地方，我们在此进行解释：

（1）关于 SRM 也利用了正误信息的批注

根据作者的理解，由于 SRM 采用了类似于 NRM 的多分类建模方式，其模型构建逻辑已经认为各状态转移之间是称名变量，没有顺序或正误之分。但如正文中所述，为了减少待估计参数数量及模型可识别问题，SRM 对区分度做出了一定的约束和限制，及设定正确状态转移的区分度为 1，错误转移的区分度为-1。但本质上，上述约束和限制并不是多分类模型本身所需的，SRM 表征的依然是多分类作答结果出现的概率，该表征并未利用行动序列的正误信息。

（2）图 2 两张图的差别是什么？

两张图没有本质上的差别，但为促进不熟悉该模型的读者的理解，我们从逻辑图和建模图两个视角阐述了该模型的建构逻辑。

（3）是否传统 IRT 模型的参数估计方法，如 EM 也可以？

由于 MCMC 算法已经满足本文所需，且其余算法是否可行并非本研究所需关注的问题。因此，我们在修改稿中没有对该问题做出相应的修改和说明。

……

审稿人 2 意见：

意见 1：对于行动序列数据，SRM 模型的目标是评价每两种状态之间进行转换的可能性，而作者开发的 ASM 模型的目标则是 评价每种可能状态转化为正确行为的可能性。从模型提供的信息来讲，两种模型各具优势；但从模型的简洁性来讲，作者开发的 ASM 更佳，具有一定的创新性。作者的研究和写作思路都比较清晰，各种对比分析都比较到位。然而，既然作者开发 ASM 模型的基本思想是借鉴 IRT 理论的 1PLM 和 2PLM 两个模型，而分步评分模型（partial-credit model,PCM）或广义分布评分模型（GPCM）同样可以用于处理行动序列数据，只不过该模型只关注每一个序列步骤正确的可能性，作者是否有分析比较 SRM，ASM 与 PCM，GPCM 两个模型之间差异与联系呢？

回应：

感谢您的审阅和肯定。需要指出，我们并没有完全理解您的意思。您提到利用 PCM 和 GPCM 处理行动序列，是否指的是将整个问题解决任务作为一道题目，依据行动序列达到任务目标的程度进行评分？以正文中图 1(a)为例来阐述我们对您问题的理解；鉴于最优问题状态转移序列(SACE)包含 3 个正确状态转移(SA、AC 和 CE)，设定该题目满分为 3 分；然后，依据被试呈现的问题状态转移序列中包含了几个正确状态转移来对其进行赋分，比如，SBDSACE = 3 分、SACSBDE = 2 分、SADE = 1 分和 SBDE = 0 分。然而，这种赋分方式有两个不足：其一，损失大量时序信息，且难以有效区分被试之间的差异。比如，SACE 和 SBDSACE 同样得 3 分，但后者在问题解决前期存在试错过程，同样的 3 分并无法有效区分它们；其二，使用这种赋分方式，一道问题解决任务只能有 1 个观测分数，而非像 ASM 和 SRM 等模型中一道问题解决任务包含多个观测分数。

此外，目前已有研究涉及 PCM 模型在过程数据中的应用(详见 3.2 节)。如，Han 和 Wilson (2022)将混合 PCM 引入过程数据分析，以及 Shu 等人(2017) 提出的 Markov-IRT 模型等。但这两个模型的分析对象都是具体的操作序列以及基于此重新编码后的具有标准化格式的数值矩阵，其中每一行代表一个被试，每一列代表一个具体操作转移，没有保留行动序列中的顺序信息。与之不同，ASM 遵循 SRM 的设定，分析的是包含顺序信息的且不具有标准化格式的状态转移序列。因此，ASM(或 SRM)和上述两个模型无法直接比较。

Han, Y., & Wilson, M. (2022). Analyzing Student Response Processes to Evaluate Success on a Technology-Based Problem-Solving Task. Applied Measurement in Education, 35(1), 33-45.

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. Psychological Test and Assessment Modeling, 59(1), 109.

意见 2：p13 页，4.2 节中作者从实测的 28851 名被试中选取了 2000 名被试，选取的方法是什么，如，随机选取；分层抽样等。

回应：

感谢您的问题，修改稿正文中已经指出“我们采用简单随机抽样，从 28,851 名被试中随机选取了 2,000 名学生…”见 4.2 节。

意见 3：4.3 节能否提供 SRM 模型参数的估计结果，从而有一个与 ASM 模型参数估计结果的直观比较。

回应：

感谢您的建议。我们在修改稿附录表 A7 中补充了 SRM 的转移倾向参数的估计结果。需要说明的是，首先，ASM 和 SRM 的模型参数含义不同，SRM 是状态之间转移的倾向性，ASM 是呈现正确行动序列的容易度。其次，ASM 和 SRM 的模型参数数量不同，如本实证研究中，SRM 包含 27 个转移倾向性参数，而 ASM 只有 10 个容易度参数。因此，两模型的模型参数(除被试能力外)估计结果没有直接可比性。

尽管两模型难以直接比较，但能从估计值中推断出相似的结论：即当被试已经处于正确问题解决路径，则其更易于保持在正确问题解决路径上；而当被试已经处于错误问题解决路径，则其更易于继续错下去，直到末尾选择乘车次数界面时才有一个纠正错误的关键期。这一点可对比参照 Han 等人(2022)的发现和本研究的发现。

Han, Y., Liu, H., & Ji, F. (2022). A sequential response model for analyzing process data on technology-based problem-solving tasks. Multivariate Behavioral Research, 57(6), 960-977.

意见 4：p17 中 logistic 回归模型的结果，建议报告模型解释的变异比是多少，即估计的能力变量能够解释数据变异的比例是多少？另外，我更好奇的是三个模型下能力参数估计值的一致性有多高，好像回归结果并不能很好的体现 这一点？从表 4 看 3 个模型的能力估计结果差异是挺大的！

回应：

感谢您的建议。首先，修改稿中我们报告了 R 方：“SRM、1P-ASM 和 2P-ASM 能力估计值的回归方程得到的 R^2 分别为 0.929、0.958 和 0.959，表明模型得到的能力估计值能够解释观测数据变异的比例很高，能够准确预测学生在任务上的作答表现，其中，2P-ASM 的变异解释率相对最大，SRM 的次之，1P-ASM 的相对最小。”见第 17 页。

其次，回归系数并非用于说明三者能力估计值之间的一致性，而是通过对问题解决结果的预测来对比三个模型的相对优劣。实证研究中，三个模型的能力参数估计值之间的一致性可通过图 7 中的散点图及相关系数来反映(比如，各模型之间估计值具有相似的分部范围和超过 0.99 的相关系数等)，而模拟研究中，三个模型的能力参数估计值之间的一致性结果呈现在表 7。

意见 5：p19，5.1 节第一段落是后一句的意思是不是指 3*2*50 实验条件下，转移状态参数都不变？如果是的话，我的理解就是 SRM 模型下的项目相关参数都不变，从而会导致 ASM 模型在每个实验条件下的 50 批数据中项目相关参数也不变，这个理解对吗？个人建议作者详细介绍一下如何利用 SRM 模型生成数据。

回应：

首先，模拟研究仅包含 $3\times2 = 6$ 个实验条件，由于参数“真值”从特定分布中随机生成，为减少随机误差对参数生成和估计带来的影响，每种实验条件下重复生成 50 组数据；在生成这 50 组数据的每一组数据时，被试的问题解决能力参数依标准正态分布重新随机生成，但状态转移参数在这 50 组数据中保持不变(仍由操纵变量的水平(序列平均长度：短或长)决定)。

另外，修改稿 5.1 节中我们进一步补充了如何基于 SRM 生成各被试的问题状态转移序列，如下：

模拟生成所有被试问题状态转移序列的具体步骤如下：

(1) 依据图 6 界定该任务的最优问题状态转移序列和所有正确/错误状态转移；

(2) 依次生成 SRM 中各模型参数，其中，

a) 被试的问题解决能力参数的“真值”依标准正态分布随机生成，$\theta_n \sim N(0,1)$；

b) 正确状态转移和错误状态转移对应的区分度参数$I_{x_j x_k}$的“真值”分别设定为 1 和 -1；

c) 状态转移倾向参数$\lambda_{x_j x_k}$的“真值”设定综合参考了实证研究中的转移倾向参数的估计值(见附录中表 A7)和 Han 等人(2022)的模拟研究设定。网络版附录中表 A3 呈现了短序列和长序列条件下所有状态转移倾向参数的“真值”；遵循 Han 等人(2022)设定，本研究中状态转移倾向参数的“真值”为固定值；

(3) 把所有参数“真值”带入 SRM，可计算得到所有被试呈现所有状态转移的概率矩阵，其中行为被试，列为状态转移；

(4) 设定所有被试从初始状态 S 开始，根据图 6 中的任务结构，在状态 S 下依据该被试呈现 SA 和 SF 的概率，使用类别分布(categorical distribution)生成第一阶段到第二阶段的状态转移(即第二阶段选择了 A 还是 F)；若选择到了 A，则在状态 A 上依据被试呈现 AB、AG 和 AS 的概率，继续使用类别分布生成第二阶段到第三阶段的状态转移(即第三阶段选择了 B、G 还是 S)；以此类推，直到抵达目标状态 J，完成该被试的问题状态转移序列生成。往复循环，生成所有被试的问题状态转移序列。

最终，本研究中生成的短问题状态转移序列和长问题状态转移序列的平均长度分别约为 10.5 和 20.2。此外，为减少随机误差影响，六种模拟条件下均按照上述数据生成步骤重复生成 50 组数据。

意见 6：附表 A3 SRM 模型项目参数的真值是如何设定或获取的，理由是什么？

回应：

对该问题的回复请见我们对您上一个问题的回复。

意见 7：作者所引用的参考文献有部分是中文出版的，建议还是在参考文献中以中文列举（原方式适合外文发表）。

回应：

本文的参考文献格式遵照了心理学报官网(https://journal.psych.ac.cn/xlxb/CN/column/column7.shtml)所要求。

第二轮

审稿人 1 意见：

意见 1：感谢作者对于审稿意见的回复，修改稿的质量得到了较好的提升。仅有以下几个细节问题供作者参考：第 8 页，18 行，“状体转移”应为“状态转移”。

回应：

感谢您的建议，修改稿中已对相关内容做出修改。

意见 2：第 9 页，11 行，“与 SRM 类似”后应有逗号分割。

回应：

感谢您的建议，修改稿中已对相关内容做出修改。

意见 3：第 15 页，15~23 行，Han 和 Wilson (2022)利用的也是状态转移，而非具体操作，Shu 等人(2017)文章中指出该模型也可以应用于状态转移。这两类模型的主要区别在于作者所列的第二点——数据格式。

回应：

感谢您的建议，我们对您指出的内容做出了修改：“比如，Han 和 Wilson (2022)将混合 Rasch 模型或混合分部评分模型应用于过程数据分析，不仅能够估计学生的潜在能力，还能够对学生的问题解决过程进行探索性分类。Shu 等人(2017)提出的马尔可夫 IRT 模型同样具有与 2P-ASM 类似的两参数 IRT 模型(或分部评分模型)形式。但上述两个模型与 ASM(以及 SRM 和状态作答模型)的主要区别在于：上述两模型分析的数据是由行动序列转化得到的具有标准化数据格式的数值型矩阵，而 ASM 分析的数据是保留了时序信息的且有个体间长度差异的非标准化格式数据。比如，前者为保证所有被试具有相同长度的数据，常把重复出现但具有前后时序的相同具体操作序列转换为频次信息并使用多级评分模型进行数据分析，但该转换损失了过程数据中重要时序信息。”

意见 4：第 17 页 21~24 行，作者给出的 R 方值“SRM、1P-ASM 和 2P-ASM 能力估计值的回归方程得到的$??^2$ 分别为 0.929、0.958 和 0.959”与描述“2P-ASM 的变异解释率相对最大，SRM 的次之，1P-ASM 的相对最小”不符，请检查。

回应：

感谢您的建议，我们在修改稿中对相关内容做了修改：“此外，SRM、1P-ASM 和 2P-ASM 能力估计值的回归方程得到的R^2分别为 0.929、0.958 和 0.959，表明模型得到能力估计值能够解释观测数据变异的比例很高，能够准确预测学生在任务上的作答表现，其中，2P-ASM 的变异解释率相对最大，1P-ASM 的次之，SRM 的相对最小。”

意见 5：第 19 页，模拟步骤编号应为（1）~（4）。 此外，文中表述“使用类别分布(categorical distribution)生成 xxx”建议修改为“根据类别分布随机生成/选择 xxx”。

回应：

感谢您的建议，后续已对模拟步骤编号进行了修改。为了使表述更为严谨，文中对您提出的内容做出了修改：“设定所有被试从初始状态 S 开始，根据图 6 中的任务结构，在状态 S 下依据该被试呈现 SA 和 SF 的概率，根据类别分布(categorical distribution)随机生成第一阶段到第二阶段的状态转移(即第二阶段选择了 A 还是 F)；若选择到了 A，则在状态 A 上依据被试呈现 AB、AG 和 AS 的概率，继续根据类别分布随机生成第二阶段到第三阶段的状态转移(即第三阶段选择了 B、G 还是 S)；以此类推，直到抵达目标状态 J，完成该被试的问题状态转移序列生成。往复循环，生成所有被试的问题状态转移序列。”

审稿人 2 意见：

意见 1：感谢作者的回复，及作者对研究的认真严谨态度。从作者修改稿可以算出作者对研究内容相关领域是比较熟悉的，但对论文本人还是想指出几个小问题：5.2 节“首先，样本量对能力参数估计的返真性的影响较小；序列平均长度越长，能力参数估计的返真性越高。”，这里需要指出的是样本量包含被试样本量和题目样本量，作者指的样本量应该是被试样本量。另外，序列的平均长度，表面是影响被试作答的序列长度，但实际反应的是题目样本量。序列平均长度越长，潜在的题目量越大。‘被试样本量越大，题目参数估计精度越高；题目样本量越大，能力估计精度越高’这一结论已成为大家的共识。从这一角度来讲，作者的结论是没有错的，但表述不够严谨。

回应：

感谢您的建议。为了使该结论更加严谨和易于理解，本文已对其进行了修改：“首先，被试样本量对能力参数估计的返真性的影响较小；序列平均长度越长，能力参数估计的返真性越高。从另外的角度来看，序列的平均长度反映了题目样本量的大小，序列平均长度越长，即题目的样本量越大，对于被试能力值的推断则越准确。”

意见 2：研究中由于 SRM 和 ASM 模型是两类不同的模型，所以作者没有比较项目参数估计的精度，是可以理解的。但两个模型是可以计算状态正确转移概率的，因此比较两类模型下状态正确转移概率的一致性是有可能的，从而可以侧面反应被试或项目参数估计是否可行。

回应：

感谢您的建议。您提到在 SRM 和 ASM 模型中比较状态正确转移概率的一致性作为评价参数估计是否可行的侧面指标。经过我们深入讨论和研究，我们认为在这个场景下进行此类比较可能并不合适。状态正确转移概率的计算方法依赖于模型结构，SRM 和 ASM 模型由于结构差异，在计算状态正确转移概率时可能会有所偏差。因此，比较这两个模型下的状态正确转移概率并不能直接反映被试或项目参数估计的一致性。

第三轮

编委意见：

该文经过修改，质量提升的同时，篇幅也变大了。建议作者将正文中可有可无的内容删减。附录可以考虑只放到网上，而不出现在印刷版中。

回应：

感谢您的意见，我们已经对文章进行了精简和调整，部分赘述内容已删除，部分次要内容（新模型与已有模型的对比和模拟研究中数据生成步骤）已放到附录，精简后正文 14000 字左右。

主编意见：

同意外审和编委意见，建议录用。