

《心理学报》审稿意见与作者回应

题目：联合作答精度和作答时间的概率态认知诊断模型

作者：田亚淑，詹沛达，王立君

第一轮

审稿人 1 意见：

意见 1：概率态 CDMs 相比于确定态 CDMs 的核心优势是什么，既是说目前已有联合作答和反应时的 CDM 了，为什么还需要概率态的 CDM 呢？第 2 节中没有阐述清楚。确定态 CDMs 得到的后验掌握概率应该也能够表征概率化的掌握情况，两者的本质区别又是什么。

回应：感谢您的问题，概率态属性比确定态属性更精细地刻画学生对属性的掌握情况，尤其是在描述学生发展时。在修改稿引言中我们做了补充。另外，在第二节中我们指出，“Zhan (2022)进一步指出 m_{nk} 和 $Prob(\alpha_{nk} = 1|\theta_n)$ 尽管有相同的表达式，但由于嵌入测量模型的方式不同，导致两者在解释上具有差异性：前者为被试对属性的掌握概率，其数值变化会直接影响被试的正确作答概率；而后者为被试归入“掌握”类的确定性程度，类似于连续变量估计中的标准误，反映的是分类的误差或确定性程度(Bradshaw & Levy, 2019)，其数值变化并不影响被试的正确作答概率，仅影响分类的确定性。”

意见 2：第二节中介绍了联合-交叉负载的建模思路（图 2 (b),图 (c)），为何指向方向是能力指向 RT 或者 α 指向 RT，而并非反向指向，这样做的依据是什么？与图 1 中其他路径指向方式相比，图 2 (b),图 (c) 这两种方式是否能为能力的估计提供更多信息？作者需阐明其中的逻辑。

回应：感谢您的问题。首先，我们未完全理解您的问题，仅根据字面理解回复。在潜变量建模中， \rightarrow 的左边通常代表的是“因”或自变量，而 \rightarrow 的右边通常代表的是“果”或因变量。通常，在潜变量模型中，潜变量为“因”而观察变量为“果”，即潜变量的变化导致观察变量的变化。在新模型中，我们认为被试的潜在能力或属性会影响观察变量 RT 的变化，因此箭头是由能力或属性指向 RT 的。

另外，交叉负载建模通过对同一潜变量增加数据量的方式理论上可以增加参数估计精度，具体而言在 RA 的基础上，通过额外增加一套 RT，实现数据量的翻倍；这类似于通过增加题目数量来提高测量精度。另外，已有交叉负载模型的研究(Bolsinova & Tijmstra, 2018; 詹沛达, 2022))以及本研究二和二的结果均表明，额外引入 RT 可以提高潜在能力参数的估计精度。

在修改稿引言中，我们指出“理论上，交叉负载建模通过成倍增加数据量来提高潜变量的估计精度(Bolsinova & Tijmstra, 2018; 詹沛达, 2022)，比如，在 RA 的基础上，通过额外增加一套匹配的 RT，实现数据量的翻倍；这类似于通过增加题目数量来提高测量精度。”

意见 3：3.3.1 数据生成部分，题目参数代表的含义，请作者描述一下，帮助读者理解。

回应：感谢建议，已添加。

意见 4：表 1 的结果呈现有点混乱，表头和数据好像不对应。

回应：感谢建议，已修改。

意见 5: 模拟研究中, 概率态属性的返真性和属性难度截距参数有关系, 文中未提及。

回应: 感谢建议, 修改稿中我们添加了简要说明: “另外, 不同属性的返真性与属性截距参数有关, 但影响不大, 基本趋势是: 属性截距参数越低(即掌握属性难度越大), 则返真性越好。”

意见 6: 实证研究中, 各个模型潜在能力估计值的散点图(图 5)呈现出的不同结果, 作者需要进行一下描述, 并对不同的现象给予一些解释, 例如联合-层级建模方式(JRT-DINA, JRT-PINC)与联合-交叉负载建模方式(C-JRT-PINC- θ , C-JRT-PINC-m)的散点图, 呈现出漏斗形, 并且在较高的能力估计值时, 散点往上倾斜。

回应: 感谢建议, 修改稿中我们添加了适当的解释: “另外, 各模型的估计值有一定差异性, 尤其是潜在能力的估计值, 这点从模拟研究中潜在能力的参数估计精度结果也可以反映出来(即两种联合建模法的主要差异体现在能力参数的估计精度上)。以 CJRT-PINC- θ 为参照基准, 对加工速度而言, 可发现(1)JRT-DINA 和 JRT-PINC 均略低估了高速度区间($\hat{\tau}_n > 1$)内被试的加工速度水平; (2)两个 CJRT-PINC 的估计结果有较高的一致性; 对潜在能力而言, 可发现(1)HO-PINC 会高估低能力区间($\hat{\theta}_n < -0.5$)内的被试; (2)JRT-DINA 会低估高能力区间($\hat{\theta}_n > 1$)内的被试; (3)两个 CJRT-PINC 的估计结果有较高的一致性; (4)对于少部分在低能力区($\hat{\theta}_n < -1$)内的被试, 基于联合-层级建模的两个模型与基于联合-交叉负载建模的两个模型之间的估计值差异较大。”

意见 7: 实证研究中出现了 JRT-DINA, 在引言部分未提及, 并且在模拟研究中也未提及。

回应: 感谢建议, 在修改稿 6.2 节中已经添加。

.....

审稿人 2 意见:

意见 1: 本研究提出的同时分析 RA 与 RT 的模型均建立在线性相关的假设上, 如 θ 与 τ 线性相关(JRT-PINC)、 θ 与 RT 线性相关(CJRT-PINC- θ)、 α 与 RT 线性相关(JRT-DINA)。在心理学研究中, 被试能力与答题时间之间的关系是存在争议的。是否有文献支持被试能力与答题速度之间的线性相关? 或者在使用本文提出的模型时, 如何检验线性相关的前提假设?

回应: 感谢您的问题。变量之间为线性相关是绝大多数以结构方差模型为例的潜变量模型的基本假设。本研究遵循了该基本假设, 且模型-数据拟合结果也表明包含该基本假设的模型可以拟合数据。

修改稿的研究局限部分我们指出: “本研究遵循联合-层级建模框架和联合-交叉负载建模框架, 假设变量之间为线性关系(如, JRT-PINC 中能力和加工速度之间, CJRT-PINC 中能力或属性和 RT 之间)。尽管本研究模型-数据拟合指标表明包含该线性假设的模型可以拟合数据, 后续研究仍可从更严谨的角度检验该线性假设是否成立; 比如, 可以通过验证能力估计值($\hat{\theta}$)与速度估计值解释后 RT 残差($\log T_{ni} - (\hat{\xi}_i - \hat{\tau}_n)$)之间是否满足线性关系来验证该假设(Bolsinova & Tijmstra, 2018)。”

意见 2: 本研究虽然从理论模型上有其合理性和价值, 但是对应用者而言, 文章显得晦涩难懂, 建议在模型提出和拓展的相应部分能够举例解释这一模型的合理性和重要性。增加文章的可读性。

回应: 感谢您的建议。该问题与问题 8 一致, 修改稿的引言和讨论中, 我们添加了相关补充说明。

意见 3: 模拟研究 1

3-1: 本研究考虑潜在能力与加工速度的相关系数($\rho\theta\tau$)为: -0.5, -0.3, 0.3 和 0.5 几个条件, 建议增加相关为 0 的条件;

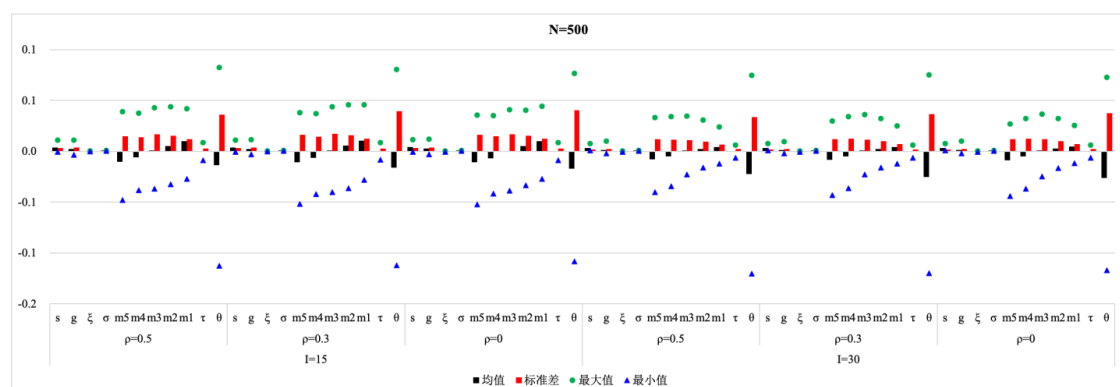
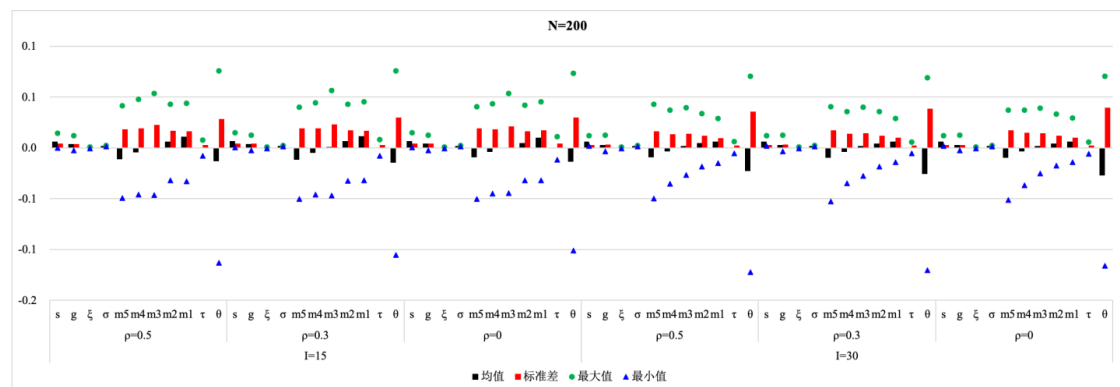
回应: 感谢您的建议, 修改稿中已经添加相关系数为 0 的情况。结果主要表明能力与速度之间的相关系数越高, 则 JRT-PINC 中潜在能力的返真性越好, 与已有研究一致。

3-2: 3.3.2 分析部分, 参数估计的复杂性对模型的应用造成了限制, 建议作者公开参数估计代码, 或整理为程序包;

回应: 感谢您的建议, 修改稿中已经将参数估计代码和事例数据上传 OSF 并提供下载链接 https://osf.io/hys7c/?view_only=cb357a6f5032424ab36b7fbda6df4d40

3-3: 参数估计值采用的是后验分布的均值还是中位数? 如果是均值, 后验分布是否为正态/对称的? 从而使均值可以较好地代表整个分布。这些在文章中缺少必要的交代, 建议补充说明。

回应: 默认采用后验均值作为报告结果, 修改中已经说明。由于 JAGS 软件主要报告后验分布中的关键值(如, 后验均值、后验中位数、后验标准差等), 并未记录完整后验分布, 所以无法以图片形式呈现后验分布。对此, 我们计算了各参数后验均值和后验中位数之差的描述性统计: 均值、标准差、最大值和最小值; 以 JRT-PINC 模型的结果为例(见下图), 结果显示两者差值较小, 表明后验分布为对称分布。



3-4: 表 1 的结果表头有误? Bias 很大, RMSE 出现负值, Cor 均较低。作者需要仔细核对结果。

回应: 已修改。

3-5: 图 4 中的结果呈现不清楚, 考虑换一种方式。另外, 不同属性估计精度存在差异, 其原因是什么?

回应: 呈现方式已经修改为表。另外, 不同属性估计精度差异是由属性难度参数设定的差异导致的, 修改稿中已经添加了相关说明。

意见 4: 模拟研究 2

4-1: 请说明生成数据所用的模型。子研究一和子研究二均需说明。

回应: 感谢建议, 修改中已经分别在子研究 1 和子研究 2 的数据生成部分添加了说明。

4-2: P3, 第 18-19 行, “与研究一中 JRT-PINC 在类似模型条件下的属性返真性相比, CJRT-PINC- θ 对属性参数的返真性优于 JRT-PINC 的(RMSE 由 0.14 左右降低 0.1 左右)”, 两个研究中数据生成的机制不同, 参数返真性的结果是否具有可比性。与上述的问题相同, P5, “整体而言, CJRT-PINC-m 在不同模拟条件下模型参数的返真性良好, 均优于不考虑 RT 的 HO-PINC; 但与 CJRT-PINC- θ 相比, CJRT-PINC-m 的参数返真性劣于前者, 表明基于交叉负载对属性(理性作答概率)引入 RT 所包含信息不如对高阶能力引入 RT 所包含信息有效。”数据生成的机制不同, 参数返真性的结果是否具有可比性。

回应: 感谢您的意见。修改稿中我们删除了研究二中的相关描述, 模型之间的对比在研究三中探讨。

意见 5: 模拟研究 3 补充: 报告数据与模型的绝对拟合情况。 P6 最后, 当生成模型为 CJRT-PINC- θ 时(相比于 CJRT-PINC-m), 两种 CJRT-PINC 的拟合优势似乎更加明显, 该如何解释?

回应: 感谢建议, 修改稿中已经添加绝对拟合结果。另外, 修改稿中, 我们在研究 3 结果部分添加了说明: “相比于 CJRT-PINC-m 作为数据生成模型, 当 CJRT-PINC- θ 作为数据生成模型时, 两个 CJRT-PINC 其相对于 JRT-PINC 的优势更明显; 这主要是由不同模型对潜在能力如何影响 RT 的假设的差异导致的, CJRT-PINC- θ 允许潜在能力对不同题目的 RT 的影响存在差异, 而 JRT-PINC 本质上是假设潜在能力对所有题目的 RT 的影响相同; 因此, CJRT-PINC- θ 可视为是 JRT-PINC 的拓展。理论上, 拓展模型可以兼容约束模型, 反之则不行。另外, 由于 CJRT-PINC-m 并没有建立潜在能力和 RT 之间的对应关系, 所以两者相互拟合各自生成的数据时受上述假设的影响较小。”

意见 6: 实证数据

6-1: 如果不考虑时间, 结果如何? 即使用 HO-PINC 分析数据, 结果与分析所用几个模型相比, 相差大吗?

回应: 感谢建议, 数据分析时额外引入 RT 的主要目的是为了提供被试的加工速度估计, 提供更丰富的反馈信息; 而提高能力或属性的估计精度是一种附带增益。修改稿中已经添加 HO-PINC 的分析结果, 与其他 3 个联合模型比 HO-PINC 的主要劣势在于无法提供加工速度的估计结果。

6-2: 请说明实证研究所用的具体参数估计方法, 基于哪种统计量计算得 ppp?

回应: 在研究三中添加了 ppp 计算, 并已提供差异测量公式。

6-3: 6.3 对应的是表 7 的结果, 而非表 12 (P8 第 10 行)。

回应: 已经修改。

6-4: P9, 第 7-10 行, “此外 CJRT-PINC-0 中各题目的交叉负载估计值之间并不一致, 表明各题目的 RT 为高阶能力提供的辅助信息量具有差异性; ”这个结果出现的原因是什么? 建议进一步解释。是否与题目的难度相关? 或者和考核的属性相关?

回应: 感谢您的建议。修改稿中我们计算了交叉负载与 3 个题目参数的相关系数, 发现交叉负载与失误参数的相关系数最高, 达到 0.923, 表明题目的失误参数越高, RT 为能力提供的辅助信息越高。当然, 这种关系只是该数据特有的, 不具有推广性。

意见 7: 上述所有研究中均未提及先验分布对参数估计的影响, 建议根据学报的要求, 补充相关信息。

回应: 感谢建议, 修改稿附录中已经通过一则模拟研究说明了模型对包含不同信息量的先验分布的鲁棒性。

意见 8: 讨论部分进一步突出本研究的实践应用价值, 对可能的应用领域和前景进行讨论。

回应: 感谢您的建议, 但限于字数限制, 我们补充的内容有限, 还请您理解。

意见 9: 文章中存在一些表述上的不规范和错别字, 建议修改。例如, 规范文章中缩写的表述, 第一次出现时使用全称+缩写, 如第 27 行; 如第 44 行, ”的到了”应为“得到了”。总体而言, 虽然在理论上具有一定的新意, 但是目前文章的可读性较差。建议修改。

回应: 感谢您的认真审阅, 我们已经通读全文修改了错别字和病句等细节性错误。需要说明的是诸如 PISA, NEAP 等常见缩写, 由于篇幅限制, 修改稿中我们仅呈现了中文名称+缩写。

第二轮

审稿人 1 意见:

意见 1: 本文新提出的方法分别释放了图 1 中对 b、c 两条路径的独立性假设, 为什么选择释放这两条假设而非其它假设(如 d、e)? 有什么合理化的依据吗?

回应: 感谢您的问题。理论上释放图 1 中任何一条独立性假设路径都可以, 但考虑到联合建模的主要目的之一是为了促进对核心特质(即潜在能力)的估计精度, 因此释放 b、c 路径是一种更常见的选择(Bolsinova & Tijmstra, 2018; 詹沛达, 2022)。修改稿中我们补充了说明: “虽然释放图 1 中任一条件独立性假设路径在理论上都是可以的, 但考虑到联合建模的主要目的之一是为了促进对核心特质(即能力)的估计精度, 所以释放 b 或 c 路径是一种更常见的选择(Bolsinova & Tijmstra, 2018; 詹沛达, 2022)。”

意见 2: 属性参数的返真性估计是否是被试掌握情况 mnk 的返真性结果? 如果是, 这种程度的偏差是否代表了模型产生了足够准确的参数估计? 建议可以做显著性检验加以证明。

回应: 是的。Bias、RMSE 和 Cor 已经从偏差、差异度和排序一致性 3 个方面反映了参数估计的返真性。显著性检验仅表明参数真值的均值与估计值的均值之间是否有显著性差异, 不足以反映参数估计精度问题。实际上, bias 反映的就是估计值的均值和真值的均值之间的差异。

意见 3: 模拟研究中返真性比拟合指标更直观准确，是主要评价指标。在模拟研究三中，缺少了对参数返真性的主要结果和结论的分析。

回应: 感谢您的建议。研究三主要目的是为了展示联合-交叉负载建模比联合-层级建模的适用范围更广，正文中的模型-数据拟合结果已经能反映该结论；因此，在字数限制的前提下我们觉得没有必要将反映相同结论的结果逐一呈现并解释。

意见 4: 确定性 CDMs 的后验概率同样可以提供被试属性掌握概率的详细信息，概率态 CDMs 也是为了表征属性的掌握概率，那两者之间的联系和区别是什么？建议作者在研究结果部分给予结果上的说明和讨论。

回应: 感谢您的建议。确定态属性的后验概率不能提供被试对属性掌握概率的信息，不应被解释为被试对属性的掌握概率。对此，文献中已有详细说明(Bradshaw, L., & Levy, R., 2019; Zhan, 2021)，包括结果上的对比(Zhan, Wang et al., 2018)。另外，在上一轮修改稿的基础上，我们进一步指出确定态属性的后验概率“为被试归入“掌握”类的确定性程度，类似于连续变量估计中的标准误，反映的是分类的误差或确定性程度(Bradshaw & Levy, 2019)；因其数值变化并不影响被试的正确作答概率，仅影响分类的确定性，所以不应被解释为被试对属性的掌握概率(Zhan, 2021)。”

意见 5: 根据以往的研究，高阶模型中中属性斜率与截距参数的估计结果往往有比较大的误差，这是否会影响对 mnk 的计算从而影响到模型的估计精度？

回应: 感谢您的建议。修改稿中，我们已经在研究局限中指出未来可操纵属性截距和区分度参数。

意见 6: 任何方法或模型均有其优势和劣势，或者是适用场景，或者是适用条件。建议作者增加对概率态 CDMs 上述这些方面的补充说明。

回应: 感谢您的建议。修改稿讨论部分已经添加少许补充说明：

“当然，本研究仅在有限的范围内探讨并展现了三个联合概率态 CDM 的表现。由于任何模型都有其适用的测验情境，本研究并没有否定其他对比模型，更多地是进一步丰富现有的认知诊断模型可选项。在实践中，我们可以尝试使用数据驱动方法，依据模型-数据相对拟合指标来选择合适的模型，进而得到客观的、准确的以及满足需要的数据分析结果。同时，本研究也启示在后续实践研究中可尝试通过多手段采集多模态测量数据，并挖掘不同模态数据所包含的信息，以期丰富研究结果并获得更完备的研究结论。”

“最后，本研究目前仅关注了 RA 和 RT 这两类数据。随着技术增强型测评的普及，多模态测评数据的采集与分析也逐渐受到人们的关注(Zhan et al., 2022)。未来可尝试对本文提出的模型做进一步拓展以期适用于更多类型数据的联合分析。”

.....

审稿人 2 意见:

意见 1: “研究一：联合-层级概率态认知诊断模型”标题的名称不完整。

回应: 感谢您的建议。根据我们的理解，我们将标题修改为“研究一：针对作答精度和时间的联合-层级概率态认知诊断模型”。类似，对研究二中的标题也做了补充。

意见 2: 3.2 “网络版附录 S1 节对比了新模型在经验信息先验下和低信息先验下的表现，两种情况下结果一致性较高，表明新模型对包含不同信息量的先验分布具有一定的鲁棒性。”这句话需谨慎，从研究结果来看，先验对于 CJRT-PINC- θ 还是有些影响的？

回应：先验分布一定会影响后验分布。我们使用了“具有一定的鲁棒性”而不是诸如“没有影响”或“完全相同”的描述，我们认为该描述是比较恰当的

意见 3：3.3.3 模拟研究中请增加关于模拟研究目的的阐述。

回应：感谢您的建议。限于字数限制，我们简单补充“本节通过一则模拟研究探究 JRT-PINC 在不同理想测试条件下的参数估计返真性。”详见第 7 页。

意见 4：4.3.3.2 “其次，JRT-PINC 的高阶能力参数和加工速度参数在所有模拟条件下的返真性均表现较好，并且速度参数的返真性明显优于能力参数。”关于速度参数返真性明显优于能力参数的结论，是因为 τ 的方差比 θ 更小，还是因为测量模型的差异？请考虑二者的可比性问题。

回应：的确如您所述，由于 RMSE 与参数真值在相同量尺上，参数真值的方差的确会影响 RMSE 的大小，Bias 也会受到类似影响。而 Cor 是从排序一致性角度反映参数估计返真性的，并不受此影响；结果显示，相同条件下加工速度的 Cor 均优于高阶能力参数的；

修改稿中我们修改了表述：“Cor 指标反映出加工速度参数的返真性优于高阶能力参数的。[脚注：由于高阶能力参数和加工速度参数的真值的方差不同(即量尺不同)，难以直接通过 Bias 和 RMSE 比较两者的返真性优劣；而 Cor 指标是从参数估计值和真值的排序一致性角度反映参数估计返真性的，并不受量尺差异的此影响。]”详见第 8 页。

意见 5：5.5.2“另外，相比于 CJRT-PINC-m 作为数据生成模型，当 CJRT-PINC- θ 作为数据生成模型时，两个 CJRT-PINC 其相对于 JRT-PINC 的优势更明显；”这句话中“两个 CJRT-PINC 其相对于 JRT-PINC 的优势更明显”表意不明，请明确表述。

回应：感谢您的建议。修改稿中我们将该内容重新表述为：“另外，当 CJRT-PINC- θ 作为数据生成模型时，两个 CJRT-PINC 的 DIC 与 JRT-PINC 的 DIC 之间的差值在 1,000 左右；而当 CJRT-PINC-m 作为数据生成模型时，两个 CJRT-PINC 的 DIC 与 JRT-PINC 的 DIC 之间的差值仅为 100 左右。该结果表明 CJRT-PINC- θ 的普适性相对更高；当其他两个模型为数据生成模型时，它能够提供更与数据生成模型几乎一致的拟合指标；而当它作为数据生成模型时，其他两个模型对数据的拟合则相对要差一些。”

意见 6：对于研究三的结果，主要用附表呈现，如“另外，各模型参数的返真性也呈现类似的趋势(见网络版附录表 S4.1-S4.4)。”但是对于研究 3 参数估计结果的解释明显显的单薄。比如表 S4.1 中，当以 CJRT-PINC- θ 为数据生成模型时，三种数据分析模型估计得到的 τ 参数的 RMSE 明显增大，该如何解释？

回应：感谢您的问题。研究三主要目的是为了展示联合-交叉负载建模比联合-层级建模的适用范围更广，正文中的模型-数据拟合结果已经能反映该结论；因此，在字数限制的前提下我们觉得没有必要将反映相同结论的结果逐一呈现并解释。

表 S4.1 中，当 CJRT-PINC- θ 为数据生成模型时，仅有另外两个不匹配的模型的 τ 的 RMSE 有所增大；与其他结果相结合，这本质上还是在反映 CJRT-PINC- θ 的普适性相对最高，这在正文中已经说明。

意见 7：最重要的问题：文章表述晦涩难懂。模型本来就复杂，一定要在文字表述上下功夫，尽量做到简洁准确，前后一致。

回应：感谢您的建议。由于本文涉及内容较多，修改稿中我们尽可能地调整了文字表达，精简掉一些赘述的内容，并且尽量保证所用名词或表述前后一致，见下表。另外，我们已经找

了2名非本专业研究生和1名本专业教授阅读本文。前两者均表示除专业术语等看不懂外，没有发现明显的病句和歧义句等，少数无法理解的地方，我们也进行了文字修改（部分修改用下划线标记）；后者表示“本文内容较多，阅读起来负担较大，但文章本身结构清晰、表述可以理解”。

原文用词	修改后用词
RA 数据、RA	RA
RT 数据、RT	RT
潜在能力、高阶能力、高阶潜在能力	能力*
潜在加工速度、潜在速度	加工速度*
认知属性、属性、潜在属性	属性*

*第一次出现时为“（潜在）属性”、“（潜在）能力”和“（潜在）加工速度”。另外，参数估计结果中为了表述准确，使用了“能力参数”、“加工速度参数”和“属性参数”。

第三轮

审稿人 1 意见：

该轮次修改基本回答了审稿人的意见，建议在字数删减时应该保留主要的研究设计、研究结论等部分，让应用者能全面认识到新模型的性能。诸如方法学上的内容可以放到在线附录部分，供建模方法的研究者详细了解即可。建议发表

回应：感谢您的肯定和建议。我们对修改稿做了进一步精简和调整，比如：将 4.2.2 节内容移至附录 S3.2。经过精简，正文（不包含表格字数）约 12,000 字。

审稿人 2 意见：

修改稿较好地解决了上次审稿发现的问题，文稿质量较前有所提升。建议对文章的语言表述作进一步修改，能在讨论部分突出本研究的实际应用价值。另外，参考文献做适当删减，格式需要进一步修改。建议修改后发表。

回应：感谢您的肯定和建议。本轮修改中，我们对文章做了进一步精简，语言表达做了微调；另外，参考文献数量由 43 篇精简至 33 篇。

编委意见：

联合作答精度和作答时间的概率态认知诊断模型一文在概率态认知诊断模型中引入题目作答时间，提出了三个概率态联合认知诊断模型，对目前这一领域的研究具有一定的拓展。论文作者对两位审稿人所提出的问题进行了系统的修改，消除了论文中的一些小的问题。建议发表。

主编意见：

经过两轮修改，稿件质量符合录用标准，同意外审和编委意见，建议发表。