《心理学报》审稿意见与作者回应

题目：模型参数点估计的可靠性：以 CDM 为例
作者：刘彦楼; 陈启山; 王一鸣; 姜晓彤

第一轮

非常感谢两位审稿专家的建设性意见和严谨评阅，以及在审稿过程中所付出的时间与精力。谢谢审稿专家对我们研究的认可。针对审稿专家所提问题，我们逐条进行了回复和修改，稿件正文中的修改部分用蓝色字体标注。

审稿专家 1 意见：

本研究关注到模型参数点估计的可靠性受到收敛判断方法、参数估计框架等因素影响，提出了提高参数点估计可靠性的新解决方法和框架，利用模拟研究验证了这个新方法和框架，得到了良好的结果，最后探讨了进一步发展的空间。

本文提出的 mCDM 模型框架采用多个初始值进行参数估计，有效优化了在非凸函数似然函数下参数估计值落入局部最优解的问题，也能解决参数估计极端值问题，使提高了参数估计的可靠性。同时，本文探讨了 4 种收敛判断方法：模型参数差的绝对值、项目正确作答概率和结构参数 组成的向量的差的绝对值、对数似然函数差、相对似然差，在此基础上提出了更加符合实际需求的综合判断法，解决了以往收敛判断标准迭代次数过少，标准适用情景单一的问题。研究结果表明，提出的收敛综合判断法 comp 在 LL 指标上也表现最佳，增强了参数估计的可靠性。最后，希望作者对以下小问题进行一定的思考:

意见 1：进一步说明一下 comp 综合收敛判断法的原理；

回应：感谢审稿人的建设性意见，我们也认为有必要进一步阐述综合收敛判断法相关原理。已在文中进行了修改，并增加了关于 comp 综合收敛判断法的原理的表述，具体内容如下（P9-P10）：

最后，阐述本文中新提出的收敛判断方法。

极大似然法估计的原理是找到最大化观察数据对数似然函数的模型参数值，并将其作为模型参数“真值”的估计。收敛判断方法的用途是判断观察数据对数似然函数的值是否已经近似达到了最大。但是，单一的判断方法在特定条件下可能存在缺陷。以对数似然函数差及模型参数差的绝对值为例进行说明。对数似然函数差方法假定第 *rep* 次及与第 *rep*+1 次迭代的对数似然函数的差小于预设的收敛容差时，似然函数值达到了最大。图 2 中呈现了对数似然函数差收敛判断方法可能存在的缺陷的简单示例。假定 B 点为 CDM 中任意一个参数的初始值$\gamma^{(0)}$。当模型参数$\gamma^{(rep)}$接近全局最优解时，如果似然函数的曲线比较平坦(可参考 Farrell & Lewandowsky, 2018)，那么将会出现模型参数差的绝对值变化较大，但是对数似然函数差变化非常小的问题。即，模型参数差的绝对值的判断效果优于对数似然函数差。模型参数差的绝对值可能存在的问题在于，似然函数值的大小除了受到模型参数值的影响之外，还受到项目数量及被试数量的影响(可参考 Rupp & van Rijn, 2018)。

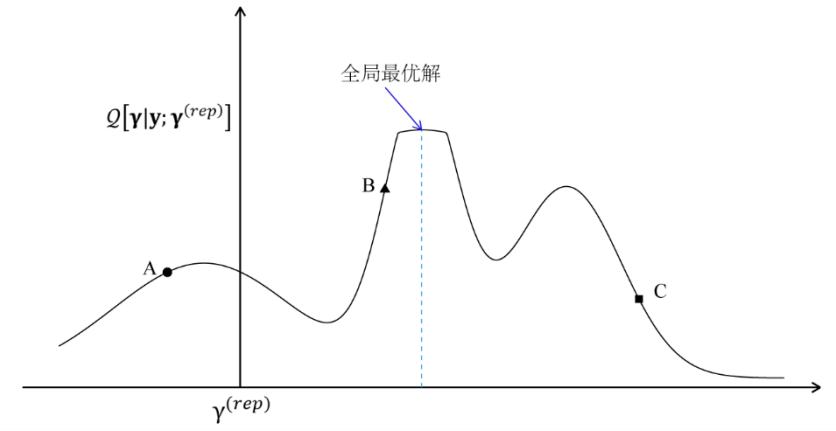

图 2 对数似然函数差收敛判断方法可能缺陷的简单示例

理论而言，进行 CDM 模型参数估计时，模型参数估计收敛判断方法及收敛容差设置越严格(这也就意味着在相同收敛容差条件下，迭代次数更多)，就越能获得使得$\ell(\hat{\boldsymbol{\gamma}}|\boldsymbol{y})$最大化的模型参数估计值。然而，实践中由于样本量、项目数量、属性数量、项目反应模型、属性层级关系及 **Q** 矩阵元素可能存在错误设定等因素的存在，很难预先判断哪种方法及相应的收敛容差是最严格的。因此，参考以往研究(George et al., 2016; Ma & de la Torre, 2020; Ma et al., 2022; Robitzsch et al., 2022; Rupp & van Rijn, 2018; von Davier, 2008; Xu & von Davier, 2008)，为克服单一判断方法可能存在的缺陷，本文提出在给定收敛容差的基础上综合使用模型参数差的绝对值、项目正确作答概率和结构参数组成的向量的差的绝对值、对数似然函数差以及相对似然差进行模型参数收敛判断，并将其称为综合判断法。

意见 2：第 21 页的 5.2 应当为 “展望”而非“讨论”。

回应：非常感谢审稿专家认真细致的建议。我们在文中对“讨论与展望”部分进行了修改，具体内容如下（P24）：

……

第三，本文新提出模型参数收敛综合判断法 comp，并在 2 种参数估计框架(*mCDM* 和 *GDINA*)、3 种收敛容差(10^{-4}、10^{-6}、10^{-8})下比较了 dp、ip、ll、rl 及 comp 等方法所组成的 30 种收敛准则的表现。就本研究所探讨的 3 种收敛容差而言，10^{-8}的表现是最好的，10^{-4}的表现则不及10^{-6}和10^{-8}；收敛容差的值越小收敛准则的表现越好，尤其是在 *mCDM* 框架下。就 dp、ip、ll、rl 及 comp 这五种收敛判断方法而言，comp 的表现最好，rl 方法的表现最差；在 *mCDM* 框架下表现最为明显。因此，本文认为，估计模型参数时，*mCDM* 框架下收敛容差为10^{-8}的 comp 方法的可靠性较高。

6.2 展望

……

审稿专家 2 意见：

研究结果的可靠性是教育和心理测量模型的重要方面，参数估计结果的可靠性与否，会影响测验结果的解释。论文《模型参数点估计的可靠性：以 CDM 为例》为提高 CDM 中模型参数点估计值的可靠性，提出了认知诊断模型（CDM）参数估计的新框架以及新的收敛判断方法；并通过模拟研究系统地探讨了新框架和新收敛判断方法的性能。研究具有较高的理论和实践价值。整体而言，该论文所研究的内容相对实用，方法合理，结论较为可靠，是一篇质量较高的学术论文。但研究还存在以下值得进一步探讨的地方：

意见 1：增加相关研究内容的文献综述，当前 CDM 点估计值可靠性方面的研究现状是什么，作者尚没有给予交代清楚。如当前是否有点估计值可靠性方面的研究，如果有，它们研究到哪个阶段了，还存在哪些不足等等。

回应：非常感谢审稿专家关于增加 CDM 点估计值可靠性方面相关研究内容的建议。我们在文中进行了如下修改（P6，P9）：

……。为提高 CDM 模型参数估计值的可靠性，研究者提出使用多个初始值(例如，300)估计模型参数(Ma & Guo, 2019)；或者是生成多个初始值(例如，200)并计算其似然函数值，然后选择似然函数值最大的那组模型参数作为 MLE-EM 迭代的初始值。图 1 中呈现了单个参数的局部最优解与全局最优解的简单示例。

……。CDM 的模型参数仅存在全局最优解的一个前提是公式**错误！未找到引用源。**为凸函数。但是，这个前提有时未必成立，导致模型参数可靠性变差。因此，参考 Ma 和 Guo (2019)的相关研究，本文提出使用多个初始值计算 CDM 模型参数。

意见 2：文献综述中各段落之间的写作逻辑有待加强，如引言第 5 段和后面 3 段之间的衔接似乎不是很紧密。第 5 段主要讲述了 CDM 的点估计算法，而第 6-8 段则阐述了 EM 算法中的初始值设置、收敛判断方法等问题。两个段落之间似乎缺少了为何要讲述初始值设置及收敛判断方法等问题的过渡语句。

回应：感谢审稿专家关于文献综述中加强各段落之间的写作逻辑的意见，我们认同专家的意见并进行了修改。对“1 引言”及“2 CDM 及其模型参数估计中存在的问题”部分的写作顺序进行了调整，加强了各部分之间的衔接。具体修改如下（P2，P6-P8）：

目前，极大似然期望最大化算法(maximum likelihood estimation using the expectation maximization algorithm, MLE-EM)是应用最广泛的 CDM 模型参数估计方法之一(de la Torre, 2009, 2011; von Davier, 2008)。例如，在 R 语言中的 *CDM* (George et al., 2016)、*GDINA* (Ma & de la Torre, 2020)软件包以及 *flexMIRT*，*Latent GOLD*，*mdltm*、*Mplus* (Sen & Terz, 2020; Templin & Hoffman, 2013)等软件中均可使用 MLE-EM 估计 CDM 的模型参数。理想条件下，使用 MLE-EM 方法能够获得具有渐近性、一致性等优良特性的点估计值。但是，研究者指出使用 MLE-EM 算法估计 CDM 模型参数时，可能会遇到的问题有：模型参数不收敛、项目参数极端值、(较差的)局部最优解以及边界值等(DeCarlo, 2011, 2019; Ma & Guo, 2019; Ma & Jiang, 2021; Philipp et al., 2018; Templin & Bradshaw, 2014; Zeng et al., 2022)。MLE-EM 估计的一般过程是，给定模型参数初始值，迭代进行 E 步(期望步)和 M 步(最大化步)，满足特定的收敛准则(convergence criterion 或 termination criterion)后停止迭代，输出模型参数的点估计值。因此，可以从参数估计框架(包括模型参数初始值设置、EM 过程等)及收敛准则等方面着手解决模型参数点估计可靠性问题。

MLE-EM 在迭代进行前需要设置模型参数初始值。CDM 模型参数估计中参数初始值向量$\boldsymbol{\gamma}^{(0)}$的设置可能会对 MLE-EM 的表现造成影响。估计模型参数时，MLE-EM 以参数初始值$\boldsymbol{\gamma}^{(0)}$为起始点通过迭代逐渐收敛到(局部)最优的模型参数估计。理想情况下，函数表达式**错误！未找到引用源。**中仅存在全局最优解，初始值$\boldsymbol{\gamma}^{(0)}$不会对最终的模型参数估计值$\hat{\boldsymbol{\gamma}}$产生影响。然而，当 CDM 的似然函数存在多个局部最优解时，初始值$\boldsymbol{\gamma}^{(0)}$不同，最终估计获得的$\hat{\boldsymbol{\gamma}}$也会不一样。即，当模型满足特定收敛准则时，模型参数估计值$\hat{\boldsymbol{\gamma}}$可能仅是一个较差的局部最优解 (Ma & Guo, 2019; Zeng et al., 2022)。……

……

收敛准则用于判断模型参数估计值是否已经足够接近模型参数最优解。一般而言，收敛准则由收敛判断方法、收敛容差及最大迭代次数这三部分组成(Paek & Cai, 2013)。收敛容差是研究者在模型参数估计前预先设定的、用于判断模型是否收敛的一个较小的值(例如，10^{-3}

或10^{-6}，甚至更小)。模型参数估计中，如果实际迭代次数没有达到预先设定的最大迭代次数，收敛判断方法在迭代前与迭代后的差异小于收敛容差，说明模型参数估计值收敛；如果实际迭代次数达到了最大迭代次数，但是收敛判断方法在迭代前与迭代后的差异没有小于收敛容差，说明模型参数估计值没有收敛，无法获得模型的极大似然估计值。

……

CDM 模型参数估计中，研究者使用的收敛判断方法、收敛容差及最大迭代次数上有明显差异。研究者经常使用的收敛判断方法是项目参数差的绝对值，且对应的收敛容差为10^{-3}或10^{-4} (参考，de la Torre 2009, 2011; Ma & de la Torre, 2016; Paulsen & Valdivia, 2022; Sen & Terzi, 2020)。一些研究者在使用项目参数差的绝对值时，将收敛容差设置的更小，例如10^{-5} (George et al., 2016)、10^{-6} (Rupp & van Rijn, 2018)或10^{-7}(Chiu et al., 2022)；也有一些研究者使用对数似然函数差进行收敛判断，并将收敛容差设置为10^{-2}或10^{-3} (Khorramdel et al., 2019; Ma & Guo, 2019)。但是 Rupp 和 van Rijn (2018)认为对数似然函数差依赖于项目数量及被试量，在进行模型参数收敛判断时相对似然差可能会更好。但是他们并没有对相对似然差的表现，以及这种方法适用的收敛容差进行研究。

意见 3：标题说的是可靠性，而正文中其实更多讲的是估计结果的可重复性问题；可靠性和可重复性是同一个概念吗？

回应：感谢审稿专家关于点估计可靠性及研究结果可重复性的问题。我们通过理论分析及统计模拟，发现使用相同模型分析相同数据时，研究结果也不一定具有可重复性。当心理统计模型中仅存在全局最优解，初始值不会对最终的模型参数估计值产生影响；但是当似然函数存在多个局部最优解时，初始值不同最终估计获得的估计值也会不一样（即点估计值不具有可重复性），进行影响到研究结论的可重复性。

极大似然法估计的原理是找到最大化观察数据对数似然函数的模型参数值。也就是，理论上而言，观察数据集相同、拟合模型相同的条件下，（根据模型参数估计值计算的）似然函数值越大，说明这组模型参数的点估计值的可靠性(可以被信赖的程度)越高。但需要特别指出的是，当似然函数存在多个局部最优解时，估计模型参数时不一定会获得全局最优解，只能获得多个局部最优解中使得似然函数相对最大的那组解。即，只能获得被信赖的程度相对较高的局部最优解，而不能保证估计值是完全可重复的。

简言之，模型参数点估计的可靠性是研究结果的可重复性的基础。

根据专家的问题，我们对正文中的有关内容进行了修改（P1，P8-P9）：

自然科学及社会科学各个领域中，研究结论的可靠性(研究结论可以被信赖的程度)，尤其是研究结果的可重复性(replication)受到极大关注(参考，胡传鹏 等, 2016; Begley & Ellis, 2012; Ioannidis, 2005, 2008; Tajika et al., 2015)。……

可以发现，研究者使用的收敛准则有很大差别。因此，相同计量模型条件下，不同的收敛准则是否会对模型参数点估计的可靠性产生影响；如果产生影响，在目前所有可用的模型参数估计收敛判断方法中，哪种效果是最好的；或者是能否开发一种具有广泛适用性的方法提高 CDM 模型参数点估计的可靠性是一个需要解决的重要问题。

……

如前所述 CDM 模型参数估计中的边界值、局部最优解、项目参数极端值、模型参数不收敛，以及收敛准则设置等可能会对模型参数点估计的可靠性产生影响，进而可能会影响到研究结果的可重复性。

意见 4：引言第二段，作者说“很少关注模型参数估计值是否可靠”，模型参数估计值的标准误不是可以作为估计值可靠性的一个指标吗？而关于标准误方面的研究，目前是存在的。

回应：假设已经研究者已经获得了可靠性较高的模型参数极大似然估计值$\hat{\boldsymbol{\gamma}}$，那么 CDM 模型参数的极大似然估计值向量$\hat{\boldsymbol{\gamma}}$与真值向量$\boldsymbol{\gamma}$的差，服从均值为 **0** 向量、方差—协方差矩阵为$\boldsymbol{\Sigma}$的多元正态分布(Liu et al., 2016)，$\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}) \xrightarrow{d} \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$。可以发现，方差—协方差矩阵为$\boldsymbol{\Sigma}$反映的是极大似然估计值向量与真值向量的差异程度。方差—协方差矩阵为$\boldsymbol{\Sigma}$对角线元素的算数平方根就是 CDM 模型参数的标准误。标准误度量的是模型参数估计的不确定性(刘彦楼, 2022; Liu et al., 2022)，而非可靠性。

获得可靠性较高的模型参数极大似然估计值$\hat{\boldsymbol{\gamma}}$，是恰当计算标准误的前提。但是，模型参数不收敛、项目参数极端值、(较差的)局部最优解以及边界值等问题，可能造成研究者无法获得可靠的$\hat{\boldsymbol{\gamma}}$。

意见 5：引言第二段，作者阐述到“使用同一个模型分析相同数据时，不同初始值可能会导致不同的模型参数估计值及 95%CI，”，审稿人觉得不同的初始值虽然会产生不同的估计值，但如果差异不大的话，基本上是可以认为结果是可靠的。有没有什么标准或者方法可用于说明确实产生了不同的参数估计值？

回应：感谢审稿专家的意见。假设使用相同模型拟合相同观察数据集时，那么观察数据似然函数值的大小仅受到模型参数估计值的影响。如果模型参数估计值近似相等的话，那么，似然函数值也应该近似相等。如果，似然函数值不近似相等，那么意味着产生了不同的参数估计值（注意，似然函数值近似相等不一定意味着模型参数估计值近似相等）。我们对原文进行了如下修改（P1）：

……。使用同一个模型分析相同数据时，不同初始值可能会导致模型参数收敛于不同的局部最优解。根据极大似然法原理，似然函数值不同，说明产生了不同的模型参数估计值；似然函数值之间的差异越大，说明模型参数局部最优解之间的差异越大。例如，假设γ是模型中任意一个参数，如果第一次的点估计值与第二次的点估计值的差$\hat{\gamma}^{(1)}-\hat{\gamma}^{(2)}$不近似为 0，说明在这两次估计中模型参数$\gamma$的估计值及 95% CI 不同。

意见 6：模拟研究中，各评价指标的计算公式未给予呈现，其指标判断标准也未说明。

回应：感谢审稿专家认真细致的建议。我们在文中增加了评价指标的计算公式，并对指标判断标准进行了说明（P12-P13）：

收敛准则的目的是判断迭代过程中的模型参数是否已经最大化了似然函数，因此本研究的评价指标主要围绕对数似然函数进行构建，包括：最佳似然函数次数($\mathrm{LL_{Best}}$)，似然函数的均值($\mathrm{LL_{mean}}$)、似然函数的最大值($\mathrm{LL_{max}}$)、似然函数的最小值($\mathrm{LL_{min}}$)以及似然函数的标准差($\mathrm{LL_{sd}}$)。最佳似然函数次数指的是 30 种收敛准则在 500 次重复中分别取得最佳似然函数值的次数$\mathrm{LL_{Best}}=\sum_{R=1}^{500} I\left(\mathrm{LL_{Conv_}}_R=\mathrm{LL_{max_}}_R\right)$；其中，$\mathrm{LL_{Conv_}}_R$表示各收敛准则在第 R 次重复中对应的对数似然函数值，$\mathrm{LL_{max_}}_R=\max\left(\mathrm{LL_{Conv_}}_R\right)$表示第 R 次重复中所有收敛准则对应的对数似然函数的最大值，I是示性函数用于判断前后两个函数值是否相等，如果$\mathrm{LL_{Conv_}}_R$与$\mathrm{LL_{max_}}_R$相等，函数I的值等于 1，否则等于 0。关于$\mathrm{LL_{Best}}$需要特别说明的是，在单次循环中可能会有多个收敛准则同时取得最佳似然函数值；$\mathrm{LL_{Best}}$的值越大说明的是收敛准则的表现越好。$\mathrm{LL_{mean}}$、$\mathrm{LL_{max}}$、$\mathrm{LL_{min}}$以及$\mathrm{LL_{sd}}$表示 500 次重复中 30 种收敛准则分别对应的对数似然函数值的均值、最大值、最小值以及标准差，例如$\mathrm{LL_{mean}}=\mathrm{mean}\left(\mathrm{LL_{Conv_}}_R\right)$。

其他评价指标还包括：500 次重复中 30 种收敛准则分别对应的模型参数估计程序单次运行的平均时间($\mathrm{t_{mean}}$，单位是秒)，平均迭代次数($\mathrm{Itr_{mean}}$)，实际迭代次数的最大值($\mathrm{Itr_{max}}$)，所有项目参数出现极端值的总数(将项目参数大于 1 或者是小于-1 定义为极端值，表示为λ_{out})，以及模型参数估计程序未收敛次数的总次数。

意见 7：缺少实证研究，建议增加实证研究，以进一步检验作者提出的新框架和收敛判断标准在实证数据中的表现情况。

回应：感谢审稿专家的建议。我们在文中增加了实证研究部分，主要修改内容如下（P1，P22-P23）：

……；第 5 部分是实证数据分析，目的是检验新提出的模型参数估计框架及收敛准则在估计 CDM 模型参数时的表现，并与 *GDINA* 软件包的表现进行比较；……

5 实证数据分析

数据来源于 Yuan 等人(2022)关于小学数学分数运算的认知诊断研究。这个数据集包含 817 名被试对 56 个项目的作答。Yuan 等人(2022)在文献分析的基础上，根据专家建议、被试访谈及口语报告法等，定义了 5 个认知属性，分别是：基本运算(α_1)、约分(α_2)、通分(α_3)、带分数拆分(α_4)、借位(α_5)。其研究提出分数运算认知过程的可能路径是：掌握α_1是掌握α_2、α_3、α_5的前提；由于属性α_4仅涉及将整数与分数部分拆开，不需要预先掌握α_1；图 4 中呈现了认知属性层级关系图。Yuan 等人(2022)使用似然比统计量比较了 logit 连接函数下饱和 CDM 与 HCDM 的对数似然函数值的差异，初步证实了小学数学分数运算数据集中存在图 4 中所呈现的层级关系。

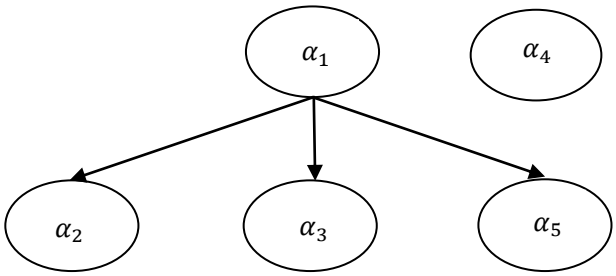

图 4 Yuan 等人(2022)定义的小学数学分数运算认知属性层级关系

本文以小学数学分数运算数据集为例，探讨当 CDM 模型中存在边界值时，*GDINA* 及 *mCDM* 模型参数估计框架下，5 种收敛判断方法(dp、ip、ll、rl、comp)，3 种收敛容差(10^{-4}、10^{-6}、10^{-8})，所组成的 30 种收敛判断准则的表现。表 7 中呈现了这 30 种收敛准则对应的对数似然函数值(简记为，LL)，以秒为单位的模型参数估计时间(t)，迭代次数以及λ_{out}；为便于结果解释，将 LL 值保留到了小数点后四位。

根据模型参数估计的极大似然理论，收敛判断准则对应的 LL 越大，说明这个准则的表现越好，模型参数点估计值的可靠性越高。

可以发现：(1)对于 LL 值影响最大的是模型参数估计框架，本研究中新开发的 *mCDM* 框架下各收敛准则对应的 LL 值远大于 *GDINA* 框架下各收敛准则对应的 LL 值。(2)在所有收敛准则中表现最好的是 mdp8 与 mcomp8，在这两种收敛准则中不仅似然函数是最大的，而且项目参数中没有极端值。(3)就 3 种收敛容差而言，不论是在 *mCDM* 还是 *GDINA* 框架下，10^{-4}的表现都是最差的，10^{-8}的表现是最佳的；尽管在一些收敛准则中10^{-6}的表现与10^{-8}类似，但是前者并不具有普遍适用性。以上 3 个发现与模拟研究中的结论具有高度的一致性。

表 7 实证数据分析结果

GDINA 框架					*mCDM* 框架				
收敛准则	LL	t	Itr	λ_{out}	Cov	LL	t	Itr	λ_{out}
Gdp4	-14307.9718	1.040	133	4	mdp4	-14248.5465	0.470	64	1
Gdp6	-14307.9717	1.328	190	4	mdp6	-14248.5463	0.718	111	1
Gdp8	-14307.9717	1.686	247	4	mdp8	-14248.5463	0.975	158	0
Gip4	-14307.9719	0.914	123	4	mip4	-14248.5469	0.423	58	0
Gip6	-14307.9717	1.299	181	4	mip6	-14248.5463	0.670	105	1
Gip8	-14307.9717	1.631	238	4	mip8	-14248.5463	0.925	152	1
Gll4	-14307.9720	0.891	119	4	mll4	-14248.5465	0.449	63	3
Gll6	-14307.9717	1.128	148	4	mll6	-14248.5463	0.570	87	1
Gll8	-14307.9717	1.245	177	4	mll8	-14248.5463	0.698	110	2
Grl4	-14351.6261	0.264	20	4	mrl4	-14257.7213	0.168	13	0
Grl6	-14308.0450	0.448	47	4	mrl6	-14248.6033	0.289	35	1
Grl8	-14307.9725	0.856	111	4	mrl8	-14248.5469	0.415	58	0
Gcomp4	-14307.9718	1.040	133	4	mcomp4	-14248.5465	0.470	64	1
Gcomp6	-14307.9717	1.328	190	4	mcomp6	-14248.5463	0.718	111	1
Gcomp8	-14307.9717	1.686	247	4	mcomp8	-14248.5463	0.975	158	0

第二轮

审稿专家 2 意见：

论文《模型参数点估计的可靠性：以 CDM 为例》经过作者第一轮修改之后，质量得到较大提升，作者合理而科学地回复了审稿人提的意见。经过再次评审，审稿人还发现以下几个问题想和作者进行探讨：

回应：感谢审稿专家对我们工作的肯定，也再次感谢专家的严谨与细致。以下是对专家意见的点对点回应，稿件正文中的修改部分用橙色字体标注。

意见 1：第三章节处理边界值问题时，作者说当公式（8）的分母小于 0.1 时，在分母上增加10^{-14}，作者可否阐述一下增加该数值的依据，即究竟多大的值才叫足够小？是否有一个下界，即增加的这个值必须小于多少？

回应：CDM 的结构参数存在边界值时，公式(8)中可能会出现$n_l^{(rep)}$近似等于 0 的情况，造成迭代异常中止。为了保证 EM 迭代正常进行，需要保证公式(8)中的分母大于 0。我们认为当公式(8)的分母的值小于10^{-16}时，已经足够小了。另外，当公式(8)中的分母的值小于10^{-16}时，分子（第 l 种属性掌握模式下正确作答项目 j 的期望人数）$r_{lj}^{(rep)}$也会小于10^{-16}。因此，为保证$P_{lj}^{(rep+1)}$取得一个较小的值(例如，小于 0.01)，我们采用的是：分母小于10^{-16}时，在分母上加上10^{-14}。根据审稿专家的问题，我们在正文中增加了如下内容（P9）：

……。第 l 种属性掌握模式下，正确作答项目 j 的期望人数(分子)不大于这个属性掌握模式下的期望人数(分母)，所以使用这个方法可以保证公式(8)中$P_{lj}^{(rep+1)}$的最大值不会超过

0.01。即，这个方法在保证分母不等于 0 的前提下，尽量减小校正系数对正确作答概率的影响。感兴趣的读者可以尝试使用其他值，但是我们认为只要满足分母不等于 0，且$P_{lj}^{(rep+1)}$较小(如，小于 0.01)这两个条件，不同的校正系数对模型参数估计结果不会产生明显影响。

意见 2：第三章节标题，作者提出了新的模型参数估计框架，可否请作者详细阐述一下哪些内容是属于新估计框架内的，哪些又独立于新估计框架？从章节标题来看，收敛标准独立于新估计框架，收敛准则为何是属于估计框架外的？

回应：感谢审稿专家提出的关于新估计框架以及收敛准则的问题。我们在正文中增加了两部分内容（P9，P11）综合阐述新估计框架内容，以及收敛准则为何属于估计框架外：

……。因此，本文提出新的模型参数估计框架试图解决 2.2 部分提及的模型参数估计中可能存在的问题；提出新的收敛准则试图解决 2.3 部分提及的收敛准则可能存在的问题。

……

综上所述，本研究提出了基于 MLE-EM 的 CDM 模型参数估计新框架及新收敛准则，以提高模型参数点估计的可靠性。新的模型参数估计框架包括对 MLE-EM 方法中的 E 步及 M 步的改进。对 E 步的主要改进是，必要时(如，模型参数不收敛或项目参数存在极端值时)使用不同的初始值分别重新计算 E 步中的期望次数及进行后续的迭代。对 M 步的主要改进是，保证公式(8)中分母不等于 0 且$P_{lj}^{(rep+1)}$较小。

意见 3：模拟研究中，作者在计算最佳似然函数次数时，无论是 LLconv_R 还是 LLmax_R 的计算，使用的全部是估计得到的似然值；这意味着 LLmax_R 也可能不是最佳的结果，而基于该非最佳似然得到的结果可能有偏。是否可以用数据生成时的模型参数来估计似然函数值（记为 LLtrue），然后将 30 种收敛准则下得到的似然函数值和 LLtrue 之间的差值作为评定标准，具有最小差值所对应的收敛准则记为最佳收敛准则？并且认定最小差值超过某个数（如 10）时，则所有收敛准则表现均比较差？

回应：根据极大似然估计的原理，MLE-EM 方法的主要目的是计算出能够最大化似然函数的模型参数值。心理统计与测量模型只是对现实世界的抽象与概括（特别说明，其他学科中也是类似的），这也就意味着现实中模型与数据难以做到完美拟合，只能找到较好地拟合观察数据的模型。正是因为以上两点原因，为了使得研究结果更好地应用于实践，本文中使用估计得到的似然值。

意见 4：实证研究中，Yuan 等（2022）已经证明该批数据所考察的属性之间存在层级关系，这意味着无论是使用 GDINA 模型还是 mCDM，其拟合结果很可能没有 HCDM 的拟合结果好。基于此，作者可否增加 HCDM 下的似然函数估计结果，以明确 mCDM、GDINA 二种估计框架下的结果估计与 HCDM 估计结果之间的差距。

回应：心理学研究中，属性之间的层级关系至少能够部分地反映出被试的认知过程，因此，对观察数据中可能存在的层级关系进行探索或验证具有重要的理论及实践价值。

目前，研究中开发的属性层级关系探索或者验证方法，据我们所知，大多数都需要饱和模型的参与（即，模型中可能存在边界值）。例如，LR 方法比较的是饱和 CDM 与 HCDM 的似然函数 (Templin & Bradshaw, 2014)，z 统计量检验的是饱和模型中的结构参数是否近似等于 0 (Liu et al., 2022)，PEM 方法是探索饱和模型中的一些结构参数能否收缩到 0 (Gu & Xu 2019;)。

本研究的主要目的之一是探讨 CDM 中存在边界值时，如何有效提高模型参数点估计的

可靠性。我们认为研究者可以进一步在本研究中新提出的 *mCDM* 框架下探讨饱和 CDM 与 HCDM 比较的问题、或者是对属性层级关系进行探索或验证的问题。根据专家的问题，结合**意见 5**，我们在讨论部分增加了关于研究者在 *mCDM* 框架下恰当使用 HCDM 分析数据的建议（P24）。

意见 5：建议作者在讨论部分增加关于实证研究的一个注意事项，即当有充分证据证明属性间存在层级关系时，最好使用层级 CDM 对数据进行分析，此处未使用 HCDM 进行分析，是为探讨存在边界问题情况下，比较不同估计框架对数据的拟合结果。

回应：感谢审稿专家的建议。结合专家的意见 4，我们在讨论部分增加了以新提出的 *mCDM* 框架为基础，如何恰当使用 HCDM，进一步提高研究模型参数估计值可靠性的建议（P24）：

……。导致 CDM 中存在边界值的一个原因是属性间存在层级关系，使得饱和 CDM 中的一些参数近似等于 0。研究者以饱和 CDM 为基础开发了一些属性层级关系探索或验证的方法(Gu & Xu 2019; Liu et al., 2022; Templin & Bradshaw, 2014)。我们建议研究者进一步在 *mCDM* 框架下使用已有方法或者是开发新方法对属性层级关系进行研究。当有较为充分的证据证明层级关系存在时，在 *mCDM* 框架下使用 HCDM 分析数据，可能会提高模型参数点估计值的可靠性。

编委专家意见：

这篇论文作者为提高 CDM 中模型参数点估计值的可靠性，提出了认知诊断模型（CDM）参数估计的新框架以及新的收敛判断方法；并通过模拟研究系统地探讨了新框架和新收敛判断方法的性能。研究具有较高的理论和实践价值。经过两位审稿人的挑剔性审阅和作者的修改，论文达到了学报发表的质量要求，建议发表该论文。

主编意见：

同意外审和编委意见，建议录用。