《心理学报》审稿意见与作者回应

题目：《一种简单有效的 Q 矩阵修正新方法》
作者：李佳；毛秀珍；韦嘉

第一轮

审稿人 1 意见：

意见 1：

“引言”部分的问题。（1）第 4 页第 4~5 行“MLE 和 MMLE 运用迭代的 EM 算法…”。MLE 方法并不使用迭代的 EM 算法。（2）第 4 页第 2 段第 2 行“…也不需要复杂的参数估计和耗时繁琐的计算”。S 统计量方法和残差方法都属于基于模型数据拟合的方法，但上一段提到前者的计算复杂，后者的过程繁琐。与这句话的描述相悖。

回应：

答：感谢专家提出的建议，作者针对上述问题分别回答如下：

（1）Wang 等人(2018)将 MLE 和 MMLE 用于 *Q* 矩阵修正，并用多次 EM 算法对 *Q* 矩阵进行重复修正。原稿中“使用迭代的 EM 算法”表述不准确，已将其修改为“其中，MLE 和 MMLE 采用 EM 算法对 *Q* 矩阵进行重复修正，修正率较高但往往比较耗时，而基于贝叶斯方法的过程复杂且易受先验分布的影响。”。修改稿的对应部分已用红色字体作了标注。

（2）基于模型数据拟合的大部分方法计算较简单，S 统计量方法和残差方法属于其中较为复杂的方法。原稿表述不完整，导致前后描述相悖。修改稿对引言部分进行了重大修改，详见引言中的红色字体部分。

意见 2：

“Q 矩阵修正方法”部分的问题。（1）第 5 页第 3 段，因为从此开始的描述都是基于“被试在项目 j 上的反应”，所以接下来所有的相关符号和公式在定义时都应该与 j 有关（比如，体现在下标中）。（2）第 6 页第 4 行“在无失误无猜测的条件下，…”。这句话是否适用于所有的认知诊断模型吗？（3）第 6 页第 5~7 行“…则期望 f_(1,1)越大…”。与什么相比越大，描述中并没有体现出比较级。（4）第 6 页第 8 行。理论上，基尼系数的值可以等于 0 吗？（5）第 8 页第 1 段最后一行。值为 2.285？（6）文中所有表格在绘制时请使用三线表。（7）公式（2）下面 1~2 行。如果后验概率是根据贝叶斯公式计算得到的，那么有没有考虑先验概率？怎么选取？（8）“2.3 Q 矩阵修正的步骤”的算法描述有待改进。本质上，实施第 4

步也需要重复执行前 3 个步骤。另外，第 4 步“…c 种可能的…”中的 c 是小 c 还是大 c？（9）“2.3 Q 矩阵修正的步骤”最后一段。理论上讲，循环修正得到的结果不应该更稳定、更稳健吗？循环修正过程中会产生不收敛的情况吗？如果会，原因是什么？

回应：

答：感谢专家认真细致的审阅，作者针对上述问题分别回答如下：

（1）原稿中多处公式和符号存在不对等的情况，修改稿已逐一进行检查修正。

（2）研究将理想反应定义为：在无失误无猜测的条件下，当被试掌握了项目考察的所有属性，则其理想反应为 1，否则为 0。该定义适用于任何认知诊断模型。

（3）文中是将 $f_{(1,1)}$ 与 $f_{(0,0)}$、 $f_{(0,1)}$ 和 $f_{(1,0)}$ 相比。初稿中表述不完整。修改稿将其修改为“若被试 $\boldsymbol{\alpha}_l$ 的理想反应为 1 且 q 向量正确时，则观察反应为 1 的被试中预测反应为 1 的人数比例也越高，即期望 $f_{(1,1)}$ 越大， $f_{(0,0)}$、 $f_{(0,1)}$ 和 $f_{(1,0)}$ 越小，四种反应类别的纯度越高；当被试 $\boldsymbol{\alpha}_l$ 的理想反应为 0 且 q 向量正确时，则观察反应为 0 的被试中预测反应为 0 的人数比例也越高，即期望 $f_{(0,0)}$ 越大， $f_{(1,1)}$、 $f_{(0,1)}$ 和 $f_{(1,0)}$ 越小，四种反应类别的纯度也越高。”。修改稿的对应部分已用红色字体作了标注。

（4）理论上，基尼系数的值不能等于 0。因为 $f_{(1,1)}$、 $f_{(0,0)}$、 $f_{(0,1)}$ 和 $f_{(1,0)}$ 的值均大于等于 0，且至少有一项的值大于 0，所以基尼系数的值恒大于 0。

（5）最终基尼系数的值是将表 4 中最后一列 $Gini_{\boldsymbol{\alpha}_l}$ 在四种知识状态下的的值相加，即 0.444+0.648+0.657+0.536=2.285。文中表 4-6 的呈现方式不清楚，作者已将其合并为一个表格，并更改了呈现方式。

（6）修改稿已将所有的表格更改为三线表。

（7）后验概率的计算需要用到先验概率，初稿中没有说清楚先验概率的选取。修改稿进行了补充说明。原稿的内容已修改为“其中， $w(\boldsymbol{\alpha}_l)$ 表示总体中 $\boldsymbol{\alpha}_l$ 的后验概率，根据 KS 的先验分布和 $\boldsymbol{\alpha}_l$ 类被试的似然计算而得。本文假设 KS 服从均匀分布。”

（8）初稿对 Q 矩阵修正的第四步描述不太完整，修改稿进行补充说明。对应部分已经修改为“第四，重复第一步到第三步，计算项目 j 在所有候选 $\boldsymbol{q}$ 向量下的 ORDP、R、RMSEA 或 HD 的值；第五，从项目 j 的 C 种可能的属性模式中选择使 ORDP、R、RMSEA 或 HD

最小的$\boldsymbol{q}_{jc}$作为项目 j 的q向量；第六，重复上述五个步骤，直到修正完所有 L 个项目，算法停止。”。修改稿的对应部分已用红色字体作了标注。

（9）感谢专家指出的问题。理论上讲，循环修正得到的结果更稳定、更稳健。例如，在 K=5, L=30, N=1000, M=20%, Iq~U[0.05, 0.25]，被试采用均匀分布，层级结构为独立型时，Yu 和 Cheng(2019)采用循环修正得到的 R 方法的 PMR 值为 0.995，本研究采用一次修正得到的 R 方法的 PMR 值为 0.968。由此可见，循环修正与一次修正的结果相比更好，但没有明显的差异。

事实上，循环修正有很多不足之处。首先，循环修正非常费时。以 ORDP 方法为例，在属性层级结构为独立型，被试知识状态为均匀分布，L=20，N=300，M=20%，Iq~U[0.05, 0.25]的条件下，循环修正需要 147s，而一次修正仅需要 12s。当被试或项目增加时，循环修正耗时会更长。其次，循环修正可能存在前后两次修正的 Q 矩阵始终不相同即不收敛的情况。这时需要设置最大迭代次数。再次，有许多 Q 矩阵修正的研究也采用了一次修正(或估计)进行实验，例如 Wang 等人(2020)、Kang 等人(2019)和涂冬波(2012)。

因此，一方面借鉴已有研究的方法，另一方面考虑到本研究的实验条件较多，同时考虑到若所有方法都采用一次修正则方法间比较的基础也相同，研究采用了一次修正 Q 矩阵的方法。修改稿在讨论部分进行了补充说明。

意见 3：

“模拟研究”部分的问题 （1）“Iq=[0.05, 0.25], [0.05, 0.4]”没有交待清楚，两个范围是什么模型什么参数的取值范围？ （2）“3.4 结果”第 2 段第 1 行的描述不够准确，“ORDP 方法在绝大多数实验条件下…”可能更为准确。（3）部分结果并不随着样本量的增加而变好（比如，表 7 中“Iq=(0.05, 0.4)、独立、L=30”条件下 ORDP 方法的两个 TPR 值【0.897 和 0.826】），可能会是什么原因？ （4）第 17 页第三点中第 3~4 行的 4 个值是怎么得到的？

回应：

答：感谢专家提出的建议和指出的问题，作者针对上述问题的回答如下：

（1）“Iq=[0.05, 0.25], [0.05, 0.4]”是 DINA 模型的失误和猜测参数的取值范围，原稿没有交代清楚，表达也有不准确的地方。修改稿已将对应部分修改为：“DINA 模型的失误 s 和猜测 g 参数均服从均匀分布。其中，高、低质量项目的 s 和 g 参数分别从区间(0.05,0.25)和(0.05,0.4)中随机产生。”

（2）在结果部分，作者表述“ORDP 方法几乎在所有实验条件下都具有最高 PMR 和

TPR 值”不够严谨，修改稿已将其修改为“ORDP 方法在绝大多数实验条件下都具有最高 PMR 和 TPR 值”，并用红色字体标注。

（3）作者查阅了做实验时保留的数据结果文件，发现此处是作者在誊抄时出现的错误，将两个数据写反了。感谢专家的仔细分析，指明错误。我们已在修改稿进行更正。

（4）由于研究的实验条件较多，以样本量 *N* 为例进行说明。首先，分别求出 *N*=300 和 1000 时所有实验条件下 PMR 的均值。然后，将两个均值相减得到仅变化样本量 *N* 时，PMR 均值的全距。初稿中对此处的描述不清楚。修改稿进行补充不说。具体内容为“固定 *N*=300(1000)时，ORDP 方法在所有实验条件下 PMR 均值为 0.905(0.930)。此时，PMR 均值的全距为 0.025。同理，固定 *N*=300(1000)时，R、RMSEA 和 HD 方法在所有实验条件下 PMR 均值的全距分别为 0.038、0.029 和 0.019。类似的，仅固定 *L*、*M* 或 *Iq* 时，各方法在所有实验条件下 PMR 均值的全距分别记为 ORDP(0.065; 0.081; 0.073)、R(0.075; 0.075; 0.122)、RMSEA(0.078; 0.098; 0.064)、HD(0.015; 0.071; 0.095)。”。修改稿的对应部分已用红色字体作了标注。

……

审稿人 2 意见：

意见 1：

引言第一段内容，以过程性评价来结合 CDT 的重要性的逻辑是什么？终结性评价仍然可以使用 CDT。

回应：

答：感谢专家指出的问题。不论是过程性评价还是终结性评价，CDT 都能提供细粒度、多维度的评估结果，具有重要实践意义。与 CTT 和单维 IRT 相比，CDT 的重要性还体现在过程性评价中为补救教学指明具体方向。为突出 CDT“细粒度”评估的重要性和特殊性，文稿在引言部分更多强调了其在过程性评价中的作用。根据专家的建议，修改稿做了适当修改。具体内容为“认知诊断理论(cognitive diagnostic theory, CDT)运用认知心理学知识分析考生的认知过程、加工技能和知识结构，并结合现代测量学知识进行诊断分析，能够提供细粒度、多维度的评估结果，适应“过程性评价”的要求，具有重要研究与实践价值。”。修改稿的对应部分已用红色字体作了标注。

意见 2：

作者提到：基于最优项目区分度视角提出的方法对 Q 矩阵修正率不高，现有的这些方法在有些条件下的修正率是不错的，作者阐述时需要注意不要太绝对，以误导读者。

回应：

答：感谢专家指出的问题。结合专家的下一条意见，我们对这部分内容作了重大修改，并用红色字体进行标注，详见修改稿引言部分。

意见 3：

引言部分阐述的不够充分，仅仅是罗列了当前对于 Q 矩阵修订的研究，对于这些方法的逻辑，延续，不同，优势与劣势等均没有介绍清楚。其次，Q 矩阵修订工作对于 CDT 的重要性的阐述也要加强，研究者/读者为何要关注这项工作？该论文可能参考了很多中文论文的写作方法，但其实这样的写法并不是很好，心理测评的文章同样需要写得引人入胜，才能扩大该领域研究的影响力。

回应：

答：感谢专家对引言部分写作方法的建议和对写作上的高标准严要求，对我们今后的写作也有很重要的指导意义。初稿一方面没有完全体现 *Q* 矩阵修订对 CDT 的重要性，另一方面对已有方法间的逻辑，延续，不同，优势与劣势等也阐述不清楚。根据专家的建议，我们对引言进行了重大修改，但是受限于自身的写作能力与分析理解能力，还请专家进一步批评指正。具体修改内容参见修改稿引言中红色字体部分。

意见 4：

Z 表示项目可能的得分值，审稿人不太理解。Z 和 y 以及 eta 有什么区别？请作者举例说明。

回应：

答：感谢专家对文稿细致的审阅。以二级计分为例，z 表示项目可能的得分值，则 z 的取值为 0 和 1。y_{ij} 与 η_{ij} 分别表示被试 i 在项目 j 上的观察反应(即实际得分)和理想反应。理想反应表示在无失误无猜测的条件下，当被试掌握了项目考察的所有属性时，其理想反应为 1，反之为 0。

意见 5：

第二部分开始段写到：前者的核心在于构建反映观察反应和期望反应的差异性或一致性指标。这句话指代不明确。

回应：

答：感谢专家指出的问题。初稿中在此处的指代不明确，修改稿将其修改为“其中，基于绝对拟合指标的方法的核心在于构建反映观察反应和期望反应的差异性或一致性指标”，并用红色字体标注。

意见 6：

审稿人根据稿件中给出的 ORDP 方法的公式，对 Gini 系数进行了化简，得到公式如下：

$$Gini_{\alpha_{lu}} = 1-(1+2s_{q_{jc}}^2-2s_{q_{jc}})(1+2(\frac{r_{lu}}{N_{lu}})^2-2\frac{r_{lu}}{N_{lu}}) \quad （A）$$

$$Gini_{\alpha_{lv}} = 1-(1+2g_{q_{jc}}^2-2g_{q_{jc}})(1+2(\frac{r_{lv}}{N_{lv}})^2-2\frac{r_{lv}}{N_{lv}}) \quad （B）$$

可以发现，当$s_{q_{jc}}=0.5$以及$\frac{r_{lu}}{N_{lu}}=0.5$时，$Gini_{\alpha_{lu}}$达到极大值 0.75；当$g_{q_{jc}}=0.5$以及$\frac{r_{lv}}{N_{lv}}=0.5$时，$Gini_{\alpha_{lv}}$达到极大值 0.75。理论上题目参数距离 0.5 越远，以及$\frac{r_{lu}}{N_{lu}}$或$\frac{r_{lv}}{N_{lv}}$距离 0.5 越远时，Gini 系数越小，这在题目参数小于 0.5 时符合逻辑，但当题目参数大于 0.5 后，Gini 系数随题目参数增大而减小，显然存在问题。又如，当$s_{q_{jc}}$较小时（$s_{q_{jc}}<0.5$），$\frac{r_{lu}}{N_{lu}}$理应较大（$\frac{r_{lu}}{N_{lu}}>0.5$），才使期望分布和观察分布的一致性较大，$Gini_{\alpha_{lv}}$较小，而如果此时$\frac{r_{lu}}{N_{lu}}$较小时（$\frac{r_{lu}}{N_{lu}}<0.5$），此时的一致性较小，但$Gini_{\alpha_{lv}}$同样可以很小，假设$s_{q_{jc}}=0.1$时，$\frac{r_{lu}}{N_{lu}}=0.8$或$\frac{r_{lu}}{N_{lu}}=0.2$的$Gini_{\alpha_{lu}}$都等于 0.442，但此时期望分布和观察分布的一致性有很大差异。针对以上存在的问题，请作者予以回答，并思考进一步完善的可能性。

回应：

答：感谢专家对该方法深入细致的分析和思考。我们在提出方法的时候也对 ORDP 方法的最值和取值特点有过思考，但不够深入。针对专家的问题，下面从三个方面和专家交流一下我们的想法。

首先，在专家的指点下，我们也对 DINA 模型下两类被试的 Gini 公式进行化简，得到与 A 和 B 相同的表达。于是，修改稿中补充了简化后的公式。

其次，当项目的属性考察模式是候选向量q_{jc}时，影响 Gini 指数的值的因素。Gini 指数是所有知识状态类的 Gini 指数之和，而每一类知识状态的 Gini 指数可通过 A 或 B 来表示。因此，“基于候选向量q_{jc}估计的项目参数$s_{q_{jc}}$和$g_{q_{jc}}$”和“依据候选向量q_{jc}将被试分为的掌握和未掌握类的情况”共同决定了项目在候选向量q_{jc}下 Gini 指数的值。其中，$s_{q_{jc}}$和$g_{q_{jc}}$

是观察反应最佳拟合q_{jc}的情况下的取值。而$\frac{r_{lu}}{N_{lu}}$依据错误Q矩阵和观察反应而来，不随候选q向量的变化而改变。值得注意的是，候选q向量不同，掌握组和未掌握组的分类也不相同。因此，我们认为可以称“理论上题目参数距离 0.5 越远，以及$\frac{r_{lu}}{N_{lu}}$或$\frac{r_{lv}}{N_{lv}}$距离 0.5 越远时，Gini 系数越小”。但是不能简单讲“当题目参数大于 0.5 后，Gini 系数随题目参数增大而减小”，因为没有考虑到 Gini 系数的其他因素的共同影响。

再次，Gini 指数的目的是比较所有候选q向量下 Gini 系数值的大小。一般地，项目在真实q向量下，具有最优项目参数估计值，掌握组和未掌握组的分类也符合实际情况。此时，通常有$s<0.5,\frac{r_{lu}}{N_{lu}}>0.5$，Gini 值偏小。而项目在错误候选$q$向量下，参数估计值可能很大（例如大于 0.5），掌握组和未掌握组的分类也与真实情况存在交叉。换句话说，无论项目在错误候选q向量下s与g的估计值如何，掌握组和未掌握组都可能同时包含有$\frac{r_{lu}}{N_{lu}}>0.5$和$\frac{r_{lu}}{N_{lu}}<0.5$的知识状态。此时，观察分布和期望分布有较大差异时，也可能出现某些 KS 类别的 Gini 值较小。总结起来，我们认为“在所有候选q向量对应的 Gini 值中，Gini 的值越小，q向量不一定最优；但是q向量最优，Gini 的值越小”。这就意味着，根据 Gini 值最小的原则可以有更大的概率得到项目真实的q向量，但也可能存在选择到错误q向量的情况，这也是 Gini 的模式判准率不能达到 1 的原因之一。针对这种情况，我们认为“当存在多个q向量的 Gini 值都最小时，可以进一步通过项目区分度$1-s-g$的最大值来确定最后的q向量”，从而对该方法进行完善。

意见 7：

作者给出的基尼系数公式如下：

$$Gini_{\alpha_l}=f_{(1,1)}(1-f_{(1,1)})+f_{(1,0)}(1-f_{(1,0)})+f_{(0,1)}(1-f_{(0,1)})+f_{(0,0)}(1-f_{(0,0)})$$

而在实际的计算过程中是否能存在 4 种类型的反应分布呢？被试在一次假设中只能存在有一种掌握模式状态，而一道题目也只能假设一种属性考察模式。那么在一次 Gini 系数的计算过程中，对于同一组被试，是否只能存在两种情况呢？即$f_{(1,1)}$与$f_{(0,1)}$，或$f_{(1,0)}$与

$f_{(0,0)}$。在 Gini 系数的计算公式中同时列出四种类型而不对其讨论是否合理呢？希望能在文章中做出解释。

回应：

答：感谢专家提出的问题。理想反应是指“在无失误无猜测的情况下，若被试掌握了项目考察的所有属性则理想反应为 1，反之为 0”。如果是依据观察反应和理想反应进行分类，每类知识状态的反应分布就只有专家提到的两种情况。

事实上，文中的 Gini 系数是联合观察反应和预期反应对反应情况进行分类，每一类被试都存在四个反应类别。首先，预测反应是根据认知诊断模型计算得到的。因为认知诊断模型可以分别得到预测反应为 1 和 0 的概率。于是，根据观察反应的人数总体，可以得到观察反应为 1 的被试中预测反应为 1 和 0 的人数；同时也可以得到观察反应为 0 的被试中预测反应为 1 和 0 的人数。由此，根据人数进一步得到四个反应类别的人数比例。值得注意的是，虽然对每一类知识状态都分为了四种反应类别，但是在 DINA 模型下，全体被试仅分成了掌握组和未掌握组两个类别。

修改稿在 2.1 第二段末尾进行了补充说明。

意见 8：

作者选择三种比较方法的原因主要是：它们的修正率更高且计算更简单。这个结论得出的前提是，已有研究者比较过。其次，不同研究的实验条件其实是不一样的，例如，ς^2 法等只有在大样本（2000 左右）表现才会好，而其他方法的样本量无需这么大。除了样本量条件外，还有其他条件，所以，简单的阐述“它们的修正率更高且计算更简单”是不够的。需要交代清楚选择的原因。

回应：

答：感谢专家指出的文稿中表达笼统与不严谨的问题。经过反复思考和查阅文献，修改稿进一步补充了选择这三种方法的原因。修改稿 2.2 第一段相关内容已修改为：

“研究选择将 ORDP 与 R、RMSEA 和 HD 方法进行比较的原因如下：首先，它们都属于数据绝对拟合指标。其中，ORDP、R、RMSEA、S 统计量和残差方法是基于模型数据拟合视角的绝对拟合指标；HD 方法是基于统计视角的非参数绝对拟合指标。特别地，R、RMSEA 和 HD 方法的计算比较简单。其次，方法间的比较不够。目前，仅 Yu 和 Cheng(2020) 比较了 R 和 S 统计量方法。他们的结果表明 R 方法在 DINA 模型下的修正效果优于 S 统计量方法。”

意见 9：

本研究仅用 DINA 模型进行模拟研究，但在前文提到不仅使用简约模型还是用饱和模型，这里面有逻辑矛盾。第二，DINA 模型是典型的非补偿模型，基于此得到的结果虽然表明 ORDP 方法在大多数条件下结果好，但这并不能代表可以推广。建议增加，补偿、加法、饱和模型的实验结果。

回应：

答：首先，原稿提到“ORDP 方法本身不受模型限制，适用于简约模型和饱和模型”。换句话说，无论使用哪个还是哪些认知诊断模型，都可以运用该方法进行计算。目前，几乎所有关于 *Q* 矩阵修正的研究都采用了 DINA 模型开展实验，这也成为本研究选择 DINA 模型的依据。

其次，针对专家这条和第 13 条件建议，我们最后选择增加 KS 服从多元正态分布的条件开展实验，并将结果呈现在修改稿。具体原因如下：第一，重新查阅有关 *Q* 矩阵修正的研究后，我们发现所有在不同模型下比较多种方法的研究都表明 *Q* 矩阵修正方法的表现不因模型的变化有较明显的变化(Wang et al., 2018; 汪大勋 等, 2019; 汪大勋 等, 2020; Wang et al., 2020)。第二，初稿的实验条件较多。具体而言，初稿考察了两种被试人数(*N*=300, 1000)、两种测验长度(*L*=20, 30)、两种 *Q* 矩阵错误率(*M*=20%, 40%)、两种项目质量(高质量和低质量)和四种属性层级结构(独立型，直线型，收敛型和分支型)。实验采用被试内设计，一共 64 种实验条件。第三，评审专家在下面提到“有研究发现，模拟研究中 *Q* 矩阵修正方法的效果受到被试知识状态的分布的影响。”。

综上，一方面为进一步验证 KS 的分布是否影响以及和怎么影响各个方法的表现，另一方面考虑到实验条件的数量很大，重复的次数较多以及《心理学报》对模拟研究的比重的限制，我们选择增加“KS 服从多元正态分布时在 64 种不同实验条件下的实验”，并在修改稿增加了相应的结果。

意见 10：

表格采用三线表。

回应：

答：感谢专家指出的格式问题。修改稿已将所有表格改为三线表。

意见 11：

模拟研究和实证研究没有交代程序是用什么软件，自编还是使用现成代码。实证数据没

有交代使用的是什么模型，审稿人推测仍然是 DINA，那问题来了：AIC，BIC 等指标也会受到模型的影响，如果更换了其他模型结果仍是新方法好吗？第二，AIC，BIC 这些属于相对拟合指标，还需要报告绝对拟合指标。

回应：

答：感谢专家对实证研究部分指出的问题。首先，模拟研究和实证研究均采用 R 语言，并自编代码完成实验。修改稿已在 3.3 部分进行补充说明，详见 3.3 部分红色字体的内容。

其次，根据专家建议实证研究部分补充了 DINA 模型下绝对拟合指标的结果，详见文中表 10。结果表明，对于这批实证数据，各个方法修正 *Q* 矩阵后均优于原始 *Q* 矩阵的拟合结果，且它们拟合的优劣依次为：ORDP 方法、RMSEA、R 和 HD 方法，这与模拟研究的结果较为一致。

再次，根据专家的建议，我们运用 GDINA 模型对实证数据的 *Q* 矩阵进行修订和拟合比较，结果见下表 1。结果表明，一方面，根据相对拟合指标-2LL、AIC、BIC 反应的各个方法的表现优劣排序不稳定也不一致。例如，依据-2LL 各方法的排序为：HD、原始 *Q*、R、RMSEA 和 ORDP；依据 AIC 各方法的排序成为：HD、R、ORDP、原始 *Q* 和 RMSEA；若依据 BIC，各方法的排序则成为：R、ORDP、RMSEA、HD 和原始 *Q*。另一方面，绝对拟合指标的值也难以一致分辨孰优孰劣。因此，修改稿没有呈现 GDINA 在实证数据中的应用情况。

表 1 基于四种方法修正后 *Q* 矩阵的拟合指标

Q 矩阵	相对拟合指标			绝对拟合指标				
	-2LL	AIC	BIC	M_2			RMSEA	SRMSR
				M_2	*df*	*p*		
*Q*_original	6672.624	7106.620	8036.280					0.078
*Q*_ORDP	6910.985	7100.990	7507.980	299.951	25	0.001	0.143	0.113
*Q*_R	6743.096	6905.100	**7252.110**	104.700	39	0.001	0.056	0.085
*Q*_RMSEA	6894.821	7120.820	7604.930	13.207	7	0.067	0.041	0.097
*Q*_HD	**6500.980**	**6874.980**	7676.110					**0.052**

我们运用 GDINA 包进行的实证数据分析，程序运行了将近 8 个小时，结果出现上述不稳定不一致的情况。我们想，可能是因为 GDINA 属于饱和模型，参数较多，在 536 人作答 15 个项目，考察 5 个属性的条件下，参数估计精度不够高，导致出现上述在不同拟合指标下表现出不稳定不一致的结果。

意见 12：

在通常情况下，Q 矩阵修正采用循环修正的方式，直到前后两次的 Q 矩阵相同或达到

迭代次数为止才停止修正，这样使题目被充分修正。作者只对所有项目进行一次修正，不能完全反映所有方法的优劣。这种方式有无前人使用？效果差异如何？请予以文献支持。

回应：

答：感谢专家指出的问题。理论上讲，循环修正得到的结果更稳定、更稳健。例如，在 K=5, L=30, N=1000, M=20%, Iq~U[0.05, 0.25]，被试采用均匀分布，层级结构为独立型时，Yu 和 Cheng(2019)采用循环修正得到的 R 方法的 PMR 值为 0.995，本研究采用一次修正得到的 R 方法的 PMR 值为 0.968。由此可见，循环修正与一次修正的结果相比更好，但没有明显的差异。

事实上，循环修正有很多不足之处。首先，循环修正非常费时。以 ORDP 方法为例，在属性层级结构为独立型，被试知识状态为均匀分布，L=20，N=300，M=20%，Iq~U[0.05, 0.25]的条件下，循环修正需要 147s，而一次修正仅需要 12s。当被试或项目增加时，循环修正耗时会更长。其次，循环修正可能存在前后两次修正的 Q 矩阵始终不相同即不收敛的情况。这时需要设置最大迭代次数。再次，有许多 Q 矩阵修正的研究也采用了一次修正(或估计)进行实验，例如 Wang 等人(2020)、Kang 等人(2019)和涂冬波(2012)。

因此，一方面借鉴已有研究的方法，另一方面考虑到本研究的实验条件较多，同时考虑到若所有方法都采用一次修正则方法间比较的基础也相同，研究采用了一次修正 Q 矩阵的方法。修改稿在讨论部分进行了补充说明。

意见 13：

根据前人的研究，模拟研究中 Q 矩阵修正方法的效果受到不同被试知识状态的分布的影响，建议补充知识状态服从多元正态分布的实验条件，考察新方法在这种情况下是否仍具有优越性。

回应：

答：感谢专家提出的建议。我们已增加 KS 服从多元正态分布时的实验，并将结果呈现在修改稿。

意见 14：

模拟研究中，在一些情况下，RMSEA 方法的指标更佳，请作者在讨论部分针对此现象进行说明。

回应：

答：感谢专家对文稿细致认真的审阅。作者再次原稿中的测验结果进行深入分析，发现仅在项目质量较低，测验较长时，部分 RMSEA 的结果优于 ORDP 方法，但是二者的差异

很小，在 0.005 以内。修改稿在讨论部分补充了对此现象的讨论。

第二轮

审稿人 1 意见：

意见 1：

作者认真回应了审稿人提出的问题并对文章做了细致修改，修改稿的质量较原稿有较大的提升。但是，目前还有一个问题值得作者再考虑和斟酌，描述如下：作者对问题“‘2.3 Q 矩阵修正的步骤’最后一段。理论上讲，循环修正得到的结果不应该更稳定、更稳健吗？循环修正过程中会产生不收敛的情况吗？如果会，原因是什么？”的回应之一是：“Yu 和 Cheng（2019）采用循环修正得到的 R 方法的 PMR 值为 0.995，本研究采用一次修正得到的 R 方法的 PMR 值为 0.968。由此可见，循环修正与一次修正的结果相比更好，但没有明显的差异”。评论：对于 0.995 和 0.968 这两个 PMR 值，由于作者并没有做显著性检验，所以说“两者没有明显的差异”不太合适。回应之二是“循环修正非常费时。以 ORDP 方法为例，……，循环修正需要 147s，而一次修正仅需要 12s”。评论：因为 Q 矩阵修正并不像“选题”那样特别强调实时，所以为了得到更准确的修正结果，147s 完全可以接受。如作者所说，当问题更复杂时，循环修正相对一次修正的用时会长更多，但很有可能“循环修正相对一次修正的修正结果也会好更多”。回应之三是“循环修正可能存在前后两次修正的 Q 矩阵始终不相同即不收敛的情况。这时需要设置最大迭代次数。……”。评论：首先，不收敛的可能原因会是什么？其次，有些学者采用的是一次修正，但也有学者采用循环修正得到了更准确的修正结果，比如 Yu 和 Cheng（2019）。综上，审稿人还是希望能够看到循环修正的结果。

回应：

答：感谢专家指出的问题。我们拟从以下三个方面进一步探讨一次修正和循环修正的特点和差异。第一，循环修正是否会出现不收敛的情况？第二，循环修正的结果是否总比一次修正的结果更好？第三，一次修正和循环修正的结果是否存在显著差异？

为探讨上述问题，我们以独立型结构，被试 KS 为均匀分布条件为例，进行了四种 Q 矩阵修正方法在 Q 矩阵错误率（20%和 40%）、被试人数（300 和 1000）、题库质量（低和高）以及测验长度（20 和 30）一共 16 种条件下进行一次修正和循环修正的模拟实验，每个实验条件重复 30 次。

循环修正的结果见表 1。从表 1 可以看出，循环修正中，四种方法的表现从高到低依次为：ORDP、HD、RMSEA 和 R 方法，这与一次修正的结果一致。

表 1 循环修正时，各方法在独立型结构下的 PMR 值

被试分布	*M*	*L*	*N*	*Iq~U*(0.05, 0.25)				*Iq~U*(0.05, 0.4)			
				ORDP	R	RMSEA	HD	ORDP	R	RMSEA	HD
均匀分布	20%	20	300	**0.978**	0.897	0.907	0.973	**0.855**	0.688	0.713	0.845
			1000	**0.992**	0.963	0.975	0.982	**0.873**	0.727	0.773	0.867
		30	300	**0.993**	0.906	0.954	0.984	**0.884**	0.716	0.834	0.877
			1000	**1**	0.977	0.976	**1**	**0.971**	0.848	0.937	0.941
	40%	20	300	**0.848**	0.777	0.782	0.845	**0.677**	0.562	0.577	0.665
			1000	**0.936**	0.870	0.903	0.927	**0.690**	0.615	0.613	0.683
		30	300	**0.984**	0.917	0.920	0.982	**0.746**	0.606	0.696	0.739
			1000	**0.989**	0.951	0.970	**0.989**	**0.838**	0.777	0.822	0.833

对实验结果我们进行了如下分析。首先，记录了每次循环修正的迭代次数，结果见表 2。由此可知，循环修正中迭代的次数都在 5 次以下，表明循环修正不收敛的情况极少。但这并不等于循环修正一定不会出现不收敛的情况，汪大勋等人(2019)曾指出，为了防止他们提出的新方法(两阶段法)出现循环迭代，需要设置迭代终止规则。

其次，对两种方式在相同条件下的 30 次结果进行配对样本 *t* 检验，结果见表 3。表中 $M_{均值差}$表示循环修正与一次修正结果的差，负值表明一次修正的结果更好，反之循环修正的结果更好。从表 3 可知，第一，64 次实验中，78%的条件下二者均值之差的绝对值低于 0.05。所以循环修正的结果略微优于一次修正的结果，但二者的差异不明显。从 $M_{均值差}$发现，33%的一次修正的结果优于循环修正，即循环修正的结果并不总是高于一次修正的结果。这可能是若在某一次循环中出现参数估计误差较大的情况会导致 Q 矩阵修正结果不准，进而将这种误差带入到下一次修正中，导致循环修正的结果更差。第二，分析平均值差异检验结果的显著性，64 次试验中有 15 次的 P 值小于 0.05，2/3 以上的条件下两种估计方法的差异不显著。第三，cohen(1992)指出，0.2、0.5 和 0.8 分别为 d 的小、中和大效应。而表 3 中 77%的 cohen’d 值小于 0.5，说明两种修正方式的结果差异的效应不大。

表 2 独立型结构下各方法在 30 次重复实验下每次循环修正的迭代次数

方法	重复次数																													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
R	3	5	4	3	2	6	3	3	5	4	3	4	4	3	2	3	2	3	3	2	3	3	4	2	2	3	3	3	2	3
ORDP	2	3	2	4	2	3	3	3	2	3	2	4	3	3	3	3	3	3	4	2	3	2	2	2	3	3	3	4	3	2
RMSEA	2	3	3	2	6	2	2	2	2	3	2	2	2	3	2	4	3	5	2	3	3	2	3	4	5	2	2	4	2	2
HD	2	3	3	3	3	2	3	2	3	2	2	3	3	3	3	3	3	2	3	2	3	2	2	2	3	5	2	3	4	3

表 3 独立型结构下一次修正与循环修正配对 *t* 检验结果

Iq	*M*	*L*	*N*	R				ORDP				RMSEA				HD			
				M 均值差	*t*	*p*	*d*	M 均值差	*t*	*p*	*d*	M 均值差	*t*	*p*	*d*	M 均值差	*t*	*p*	*d*
Iq~*U*(0.05, 0.25)	20%	20	300	-.00038	-1.000	.326	.003	-.00070	-1.000	.326	.012	-.00045	-1.000	.326	.006	-.00113	-1.000	.326	.029
			1000	.03000	1.083	.288	.298	.03000	1.337	.192	.375	.02833	.867	.393	.231	.01500	1.795	.083	.423
		30	300	-.00090	-1.000	.326	.007	-.00110	-1.000	.326	.009	-.00112	-1.000	.326	.009	.00055	1.000	.326	.027
			1000	.00333	.067	.947	.016	-.02667	-.575	.569	.146	-.08833	-1.956	.060	.512	-.04667	-3.304	.003	.670
	40%	20	300	.12500	3.447	.002	.876	.06667	1.345	.189	.393	.06833	1.252	.221	.336	.00085	1.000	.326	.007
			1000	.07000	3.138	.004	.442	.08167	2.131	.042	.480	.08833	1.672	.105	.416	.01583	.623	.538	.144
		30	300	.10667	2.036	.051	.618	.02167	.523	.605	.132	.08333	1.670	.106	.444	.01268	.538	.595	.143
			1000	-.00500	-.103	.919	.029	-.02333	-.444	.660	.129	-.14000	-2.531	.017	.707	-.00116	-1.000	.326	.016
Iq~*U*(0.05, 0.4)	20%	20	300	-.00778	-.436	.666	.119	.00778	1.022	.315	.217	-.02556	-1.838	.076	.448	.05111	4.121	.000	.755
			1000	.03333	2.493	.019	.761	.01778	3.117	.004	.773	.00889	1.034	.310	.270	.04667	3.619	.001	.738
		30	300	-.00667	-.266	.792	.067	-.01333	-.505	.617	.144	-.03667	-1.172	.251	.327	.03556	2.372	.025	.613
			1000	.01111	.571	.573	.158	.03111	1.763	.088	.462	.01000	.625	.537	.159	.02444	1.452	.157	.438
	40%	20	300	.00011	1.000	.326	.002	.03333	4.651	.000	.989	.00778	.262	.795	.068	.00352	.259	.798	.028
			1000	.03222	2.321	.028	.626	.02000	1.079	.290	.289	.01889	.950	.350	.263	.03826	2.366	.025	.267
		30	300	-.03333	-1.020	.316	.227	.07111	2.251	.032	.513	-.08778	-2.299	.029	.591	.03556	1.510	.142	.389
			1000	.04000	.947	.351	.257	.00667	.129	.898	.036	.02667	.580	.566	.171	.07433	3.690	.001	.783

审稿人 2 意见：

意见 1：

作者在文章中，也在回复审稿意见中阐述到：“在无失误无猜测的条件下，当被试掌握了项目考察的所有属性，则其理想反应为 1，否则为 0。该定义适用于任何认知诊断模型”。审稿人认为该定义对完全补偿的 DINO 模型（只要掌握其中一个属性就可以答对），及其他非 DINA 链接的模型适用吗？如果不适用，会对 ORDP 方法在除 DINA 模型以外的模型中的表现有影响吗？

回应：

答：感谢专家认真细致的审阅。经过专家的点拨，我们讨论和思考后认为之前的定义适用于非补偿性和部分补偿性的题目。因此，对定义调整为“在无失误无猜测的条件下，当被试掌握了正确作答项目所要求的属性时，则其理想反应为 1，否则为 0。”使其适合于所有 CDM。我们已在文稿中做了修改。

ORDP 方法将用到理想反应、期望反应和观察反应。其中，期望反应由模型计算的概率来确定，随模型的不同而不同；理想反应只与被试的 KS 和正确作答项目要求的属性有关。因此，ORDP 方法适用于其它 CDM。但是，ORDP 方法的表现是否受补偿和非补偿模型的影响还值得后续的深入研究和论证。

意见 2：

研究中待估计的 Q 矩阵生成方式中的错误率，指的是 Q 矩阵中错误的项目数量占整个 Q 矩阵项目数量的比例，还是指错误的属性数量占所有属性数量的比例呢？需要阐述清楚。

回应：

答：感谢专家提出的细致思考。本文中 Q 矩阵错误率指的是 Q 矩阵中错误的项目数量占整个 Q 矩阵项目数量的比例。原稿表述不清楚，已在文章的相应部分进行修改。

意见 3：

HD 方法的逻辑本质上是符合 DINA 模型的逻辑，然而本研究中的 HD 表现最差，这是不符合逻辑的。汪大勋等人的文章（2018）中，在题目质量均为 $gs\in[0.05,0.25]$，KS 均为均匀分布时，HD 的表现仍然很不错，造成这两者差异的原因是什么呢？

本着交流讨论的目的，审稿人也进行了求证模拟实验，模型为 DINA，设置题目质量为 $gs\in[0.05,0.25]$，KS 均匀分布，Q 矩阵错误率取值 0.2，修订时不迭代，100 次循环。由于没有看明白待估计矩阵错误率的含义，所以将两种可能的方式都试了一遍，发现当错误率指错误属性数量占总体比例为 0.2 时，PMR=0.8220，当错误率指错误项目数量占总体比例为

0.2 时，PMR=0.9975，都和作者文章中的数据相差甚远。这也有可能是因为使用的 Q 矩阵不同造成的，作者能否补充一下实验中所使用的 Q 矩阵呢？

使用的 Q 矩阵共 20 题，5 属性，属性间独立，具体如下：

1 0 0 0 0

0 1 0 0 0

0 0 1 0 0

0 0 0 1 0

0 0 0 0 1

1 0 0 0 0

0 1 0 0 0

0 0 1 0 0

0 0 0 1 0

0 0 0 0 1

1 1 0 0 0

1 0 1 0 0

1 0 0 1 0

1 0 0 0 1

0 1 1 0 0

0 1 0 1 0

0 1 0 0 1

0 0 1 1 0

0 0 1 0 1

0 0 0 1 1

回应：

答：感谢专家对文章细致的把关和思考。我们对程序再次进行了仔细检查，结果发现了之前程序编写中的一个文件指代错误的问题，导致 HD 方法的结果不对。具体原因是，在编写 HD 程序的时候，按照汪大勋等人(2018)的思路，需要构造所有理想掌握模式在所有题目上的理想反应矩阵，汪大勋等人将所有理想掌握模式记为 IMP，而本研究的程序中所有理想掌握模式却为 Qs，IMP 为所有被试的知识状态。之前在程序中误将原本是 Qs 的文件写成了 IMP，导致知识状态判的不准，出现 HD 方法的修正率不高。作者重新运行了所有试验条件

下 HD 方法的结果，并在文稿中进行修正。

此外，本研究的 Q 矩阵生成方式为在保证一个 R 阵的基础上，剩下的题目从所有可能的 q 向量中随机抽取，即每个 q 向量最多考察 5 个属性。因此，本研究的 Q 矩阵与专家和汪大勋等人(2018)试验中的 Q 矩阵相比更复杂。结果发现，HD 方法在一定程度上受到 Q 矩阵影响。这和本实验中 Q 矩阵的失误率越高，修正率越低的结果相似。

第三轮

审稿人 1 意见：

意见 1：

中文摘要部分的表述还需要更精炼一些。

回应：

答：感谢专家对文章表达指出的问题。作者已精炼了中文摘要部分的表述，并在文中相应位置进行了修改。

意见 2：

第 20 页最后 1 行，q_{jc}中的 c 是从 1 到 2^(K-1)还是(2^K)-1？

回应：

答：感谢专家指出的问题。$\mathbf{q}_{jc}$ 中的$c=1,2,...,2^K-1$，这里属于作者表达错误，已在相应位置进行修改。

意见 3：

表 1 下面第 6 行，是不是漏掉了一项内容？

回应：

答：感谢专家细致的审阅。原文中“即期望 $f_{lj(1,1)}$ 越大， $f_{lj(0,0)}$、 $f_{lj(0,1)}$ 和越小”，其中“和”字后面漏掉了公式 $f_{lj(1,0)}$，现已更改。

意见 4：

表 2 下面第 1 至 2 行的这些记号是否有必要？表 3 同理。

回应：

答：感谢专家提出的意见。原文中作者写的这些记号“记 $r_{lju}\cdot(1-s_{\mathbf{q}_{jc}})/N_{lu}=A_{lju}$，$r_{lju}\cdot s_{\mathbf{q}_{jc}}/N_{lu}=B_{lju}$ $(N_{lu}-r_{lju})\cdot(1-s_{\mathbf{q}_{jc}})/N_{lu}=C_{lju}$，和 $(N_{lu}-r_{lju})\cdot s_{\mathbf{q}_{jc}}/N_{lu}=D_{lju}$。”，是

为了方便呈现公式的推导过程。考虑到行文的简洁性，作者未将公式推导过程呈现在文中，所以没有必要再呈现这些记号，现已删掉这些记号。

意见 5：

“3.1 研究目的”部分第 1 段，“研究模拟了 6 个实验变量”的表述不妥，可使用“考虑”？Iq~U[0.05,0.25]或 Iq~U[0.05, 0.4]的表述不准确，Iq 仅仅指代“题目质量”。

回应：

答：感谢专家对文章表述提出的意见。作者已修改为“研究考虑了 6 个实验变量”。并在文中将“Iq~U[0.05,0.25]或 Iq~U[0.05, 0.4]”修改为“高低 *Iq* 的参数取值范围分别为[0.05, 0.25]和[0.05, 0.4]”。

意见 6：

“3.3 评价指标”部分第 1 段，“研究自编 R 语言”的表述不妥。

回应：

答：感谢专家指出的问题。作者已在文中相应位置将“研究自编 R 语言”修改为“研究使用 R 语言程序，自编计算机代码进行模拟研究”。