《心理学报》审稿意见与作者回应

题目：基于类别水平的多级计分认知诊断 Q 矩阵修正：相对拟合统计量视角
作者：汪大勋，高旭亮，蔡艳，涂冬波

第一轮

审稿人 1 意见：

Q 矩阵修正在 CDM 中的具有重要价值，《基于类别水平的多级计分认知诊断 Q 矩阵修正》一文以 Ma 和 de la Torre（2016）提出的 seq-GDINA 模型为例探讨了多级计分的 CDM 中项目类别水平的 Q 矩阵修正方法，该研究选题具有较大的理论意义及实践价值。具体审稿意见：

意见 1：目前，在国际上已有类似研究，可在 R 软件的 GDINA 包中找到相关参考文献及实现方法。作者需要说明与类似方法相比本研究中所用方法的优势与不足。我建议通过模拟研究的方式与当前方法加以比较，但鉴于 GDINA 包中的一些方法的原理理解起来比较困难，如果作者不能实施我也能理解。

回应：谢谢专家的意见，我们也看到 Ma 和 de la Torre(2019)提出了 wald 检验和 GDI 方法相结合的方法用于 Q 矩阵修正，他们的实验也表明该方法对多级计分模型 Q 矩阵修正具有较好的效果。但如专家所言，该方法的原理理解比较复杂，在确定每个类别所测量的属性时需要进行多次的 wald 检验，并计算标准误。而本文尝试的 BIC 方法原理相对简单、计算并不复杂，且在使用新提出顺序算法（sequential search algorithm）以后大大减少了运行时间，并且模拟研究和实证研究结果显示 BIC 方法对多级计分模型 Q 矩阵的修正具有较好的效果。此外，两种方法对 Q 矩阵的修正思路并不相同，Ma 和 de la Torre 的方法是从属性的角度来检验该类别是否测量了某个属性，而 BIC 方法是从整个模型拟合的角度来比较不同 q 向量的优劣，因此两种方法是从不同角度来考察 Q 矩阵是否合理。当然未来研究可以进一步考虑与该方法甚至其它方法的比较。

意见 2：CDM 又被称为诊断分类模型，它在心理、教育等多个领域都有重要应用，作者在引言部分仅陈述它在教育测验中的价值我认为这是不妥的。

回应：谢谢专家的意见。作为新一代测量理论的核心，认知诊断对心理和教育等领域都产生

了重要作用。我们已经根据意见修改了引言部分的表述。

意见 3：在二值计分的 CDM 中有多种方法可以推广到多级计分的 CDM 中，作者为什么选择了文中的这几种，请具体阐述原因，我认为这对于读者理解本文是有帮助的。

回应：谢谢专家的意见。关于 Q 矩阵修正的研究，研究者们已经提出了许多种方法。但是部分方法具有模型限制，只能用于简化的模型，使用范围有限。少部分方法既可用于简化模型也可用于饱和模型，适用范围更广，如 GDI 法(de la Torre & Chiu, 2016)、基于似然的方法（Xu & Shang，2018）和基于残差的方法 (Chen, 2017)等。在这三种方法中，GDI 法计算相对复杂，且需要设定一个截断值（PVAF=0.95），此外我们的预研究发现该方法受样本量的影响较大。而基于残差的方法虽然可以在测验层面考察测验属性是否多余或缺失，但该方法对题目层面的属性多余不够敏感(Chen, 2017)。而 Xu 等人（Xu & Shang，2018）将基于似然的信息指标（AIC 和 BIC）用于饱和模型的 Q 矩阵修正具有较好的效果。此外，Chen，de la Torre 和 Zhang（2013）也将-2LL、AIC 和 BIC 用于对不同 Q 矩阵的鉴别。研究发现在 DINA 模型中，-2LL 指标表现较好；而在饱和模型中，BIC 指标的表现是最出色的。因此，基于以往研究，本文尝试将-2LL、AIC 和 BIC 用于多级计分模型的 Q 矩阵修正。我们在文章的引言部分重新对文献进行了梳理，并阐述了选择这三种方法进行 Q 矩阵修正的原因。

意见 4：具体研究设计部分：（1）在 CDM 中属性之间是中等或高相关的，建议作者更改属性掌握模式的产生方式，我认为这很重要。

回应：感谢专家的意见。我们这次将属性掌握模式的产生方式修改为了多元正态分布 $MVN(\mathbf{0},\boldsymbol{\Sigma})$。并根据以往研究（Liu, Xin, Andersson, & Tian, 2019；Chen，2017），将属性间的相关设置为 0.5，并更新了所有新的实验结果。

意见 5：Q 矩阵错误具体是怎么模拟的，建议作者参考 Chen, de la Torre, & Zhang, 2013；Liu, Tian, & Xin, 2016； Liu, Xin, Andersson, & Tian, 2019 等论文中的表述加以详细说明，这有助于后续研究者理解论文。

回应：参考已有研究（Chen, de la Torre, & Zhang, 2013；Liu, Tian, & Xin, 2016；Chen，2017；Liu, Xin, Andersson, & Tian, 2019）的设计，我们对错误 Q 矩阵的模拟方法进行了修改，具体模拟了 6 种类型的 Q 矩阵，包括属性冗余、属性缺失、属性既缺失又冗余，以及不同错

误比例的随机模拟。Q 矩阵错误的具体模拟方式请见文章“4.1.3”部分。

意见 6：作者在文中陈述“根据表 3 的结果，在 seq-RRUM 模型中，……MMLE 方法模式判准率在所有实验条件下均接近于 0。”，对比以往研究（Liu, Tian, & Xin, 2016），我对这个结果感到比较惊讶，请作者具体说明估计程序，并检查估计程序有无错误，如果没有错误的话请具体说明可能的原因是什么。

回应：谢谢专家的意见。本研究中的模式判准率为估计的 Q 矩阵和真实 Q 矩阵中相同 q 向量的比例。在重新进行模拟实验以后，MMLE 方法（等同于-2LL 方法）在 seq-RRUM 模型中的平均模式判准率为 15.3%。与第一稿中的模式判准率略有不同，但依旧很低，这是由于第一稿中的属性掌握模式与和这次修改中的模拟方法并不相同。此外，本研究中将 MMLE 方法（等同于-2LL 方法）用于 seq-RRUM 和 seq-GDINA 等复杂模型的 Q 矩阵修正时，会出现属性个数多的 q 向量产生的边际似然大于属性个数少的 q 向量。例如题目 j 第 h 类别的真实 q 向量为[11000]，而将[11111]作为题目 j 第 h 类别的 q 向量时，计算出的边际似然要大于真实 q 向量。这是由于在复杂模型中，属性个数越多的 q 向量具有更多的参数，因此会产生更高的似然。Chen 等人（Chen, de la Torre, & Zhang, 2013，Chen，2017）的研究中也出现了相同的结果。这也是在复杂模型中需要使用 AIC 或者 BIC 对模型参数个数进行惩罚的原因。而 MMLE 方法（或-2LL 方法）倾向于将所有类别的 q 向量定义为[11111]，所以在计算修正后 Q 矩阵与真实 Q 矩阵的模式判准率时就会特别低，也说明 MMLE 方法（或-2LL 方法）不适用于复杂模型的 Q 矩阵修正。

意见 7：建议作者增加实证数据分析部分，向研究者尤其是 CDM 的实践研究者说明及展示本研究的理论及实践价值。

回应：感谢专家的意见。我们增加了实证数据（TIMSS，Trends in International Mathematics and Science Study，2011）分析作为示例。Park, Lee 和 Johnson（2017）以及 Ma 和 de la Torre(2019)为该数据标定了 Q 矩阵，详见文章表 6。实证数据分析发现通过 BIC 方法修正后的 Q 矩阵比修正前的 Q 矩阵更加拟合实测数据。具体实证数据分析请见“研究三：实证数据分析”部分。

意见 8：文章中存在大量的表达、格式不规范等细节问题，我不一一指出了，但请作者认真修改。

回应：非常感谢专家的细致审稿。我们认真检查并修改了文中的一些表述和格式。

……

审稿人 2 意见：

这篇研究尝试提出修正基于类别水平的饱和多级计分认知诊断模型 Q 矩阵的方法。使用了 MMLE 以及在此基础上的 AIC 和 BIC 三类相对拟合指标，从一个整体拟合的角度去寻找一个使得整体拟合最好的 Q 矩阵。

意见 1：本文的最大问题是严重欠缺理论基础，进行 Q 矩阵修正的研究有很多，但本文引用的关键文献无论是 CD-CAT 还是 Chiu 的 Ideal Response Pattern 算法上都不是使用 MMLE 进行 Q 矩阵修正的，作者也没给出 MMLE 算法在饱和多级计分认知诊断模型上进行 Q 矩阵修正的任何理论推导或完备性证明。在二级计分的饱和模型上目前都不存在此类文献（只有简约模型的），因此很让人怀疑在多级计分饱和模型上 MMLE 的理论可行性和完备性。缺乏理论基础作指引，模拟研究不具有针对性和说服力，建议作者先完善相关的理论推导，再据此制定更有针对性的模拟研究。除此之外还要提供真实数据的应用例子。

回应：非常感谢专家的意见，根据专家的意见，我们对文献综述进行了重新梳理和表述，请见文中引言部分。对于使用基于似然的方法进行 Q 矩阵修正，已有研究不仅对简化模型还对饱和模型（如 LCDM 模型）进行了探索，并且从数学上进行了论证（Xu & Shang，2018）。Xu 和 Shang（2018）将基于似然的方法进行 Q 矩阵修正，并证明了基于似然的方法用于 Q 矩阵的可行性。其研究结果发现将信息指标（BIC）用于饱和模型的 Q 矩阵估计具有较好的效果。此外，Chen，de la Torre 和 Zhang（2013）也将-2LL、AIC 和 BIC 指标用于对不同 Q 矩阵的鉴别。研究发现在 DINA 模型中，-2LL 指标表现较好；而在饱和模型中，-2LL 倾向于选择在原有 Q 矩阵基础上增加属性的 Q 矩阵，而 BIC 指标的表现是最出色的。而本研究中将这三个指标用于 Q 矩阵修正的方法，实质上也是将这三个指标用于对不同 Q 矩阵进行选择，其研究逻辑与 Xu 和 Shang（2018）和 Chen 等人（Chen，de la Torre & Zhang，2013）的研究相同。

此外根据专家的意见我们重新设计了模拟研究和实证数据分析，在模拟研究部分，考虑了 6 种 Q 矩阵的错误类型，包括属性缺失、属性冗余、属性既缺失又冗余，以及随机错误，并增加了 Q 矩阵修正前后的绝对拟合指标用于评价修正后的 Q 矩阵。并通过实证数据分析来说明 BIC 方法修正 Q 矩阵的可行性。

意见 2：研究方法部分，研究应该提供修正前与修正后的 Q 矩阵在绝对拟合上的表现的对比，从而使该 Q 矩阵修正方法更具说服力。

回应：非常感谢专家的意见。根据专家的意见，我们增加了修正前后 Q 矩阵的绝对拟合指标 RMSEA（Liu, Tian, & Xin, 2016）。在三种方法中，BIC 方法对 Q 矩阵的修正效果最好，因此我们比较了修正前的 Q 矩阵和 BIC 方法修正后的 Q 矩阵的绝对拟合指标。结果显示，BIC 方法修正后的 Q 矩阵与数据更加拟合。具体结果请见文中表 3-表 5。

意见 3：Q 矩阵错误模拟部分，文章仅提到随机生成不同错误率的 Q 矩阵，并没有对 Q 矩阵存在错误的类型进行进一步的分析，如对 Q 矩阵修正研究中常探究的 over-specified，under-specified 和两者都有的情况；每个类别中的矩阵 element 改变的数量是否有限制，以及产生的错误 Q 矩阵应该满足什么要求（比如，每个类别所需要的属性个数是否有限制）等等都需要更仔细的考虑。

回应：根据专家的意见，并参考已有研究（Chen, de la Torre, & Zhang, 2013；Liu, Tian, & Xin, 2016；Chen，2017；Liu, Xin, Andersson, & Tian, 2019）的设计，我们将错误 Q 矩阵的模拟方法进行了修改，具体模拟了 6 种类型的 Q 矩阵，包括属性冗余、属性缺失、属性既缺失又冗余，以及不同错误比例的随机模拟。Q 矩阵错误的具体模拟方式请见文章“4.1.3”部分。

意见 4：同时，如果测验题目较多，属性、类别较多时，该算法步骤是否需要耗费大量的时间，研究应该考虑到方法的实际应用的可行性。

回应：感谢专家的意见，本文中的方法进行 Q 修正时确实比较耗时，且运行次数随属性个数增加而呈指数增长。因此根据专家的意见，我们对 Q 矩阵修正算法进行了优化，提出了顺序算法。顺序算法的详细介绍请见文中“3.4”部分。与原有算法（穷尽算法）相比，顺序算法对每个类别的修正不再计算所有可能的 q 向量，而运算次数会根据 q 向量的错误程度决定。相对于穷尽算法，顺序算法能大大降低运算量和运算时间。两种算法对 Q 矩阵的修正结果在大部分实验条件下一致，个别实验条件下的属性判准率差异也不超过 1%。而两种算法的运行时间上，顺序算法能大大节约 Q 矩阵修正的运行时间。再次感谢专家的意见。

第二轮

审稿人 1 意见：

意见 1：作者较好地回答了我提出的问题。并且在原有研究的基础上提出了改进的顺序算法（sequential search algorithm），增加了实证数据分析，这为本文增色不少。我有以下几个建议请作者考虑： 文中使用的 Q 矩阵修正方法都是在模型相对拟合水平上进行的修正；与

Wald 统计量相比，这些统计量只能相对的说明在项目中增加或减少某个属性是否有助于提高模型—数据的近似拟合水平，无法明确说明在项目中增加或减少某个属性是否是真正的恰当。建议作者在讨论部分增加相关表述，这关系到本研究的严谨性与创新程度。

回应：感谢专家的意见。我们已经根据专家的意见在文章讨论部分增加了相对拟合指标方法以及 stepwise 方法（Ma & de la Torre, 2019）进行 Q 矩阵修正的原理和特性。如专家所说，本文中的方法是从模型拟合的角度来判断修改某些属性是否会提高模型的整体拟合水平，而 wald 检验是从属性的角度出发验证某个属性是否对答对题目的概率有影响，两种方法对 Q 矩阵的修正结果并不能说明在项目中修改某些属性是否恰当。所有从数据角度得到的 Q 矩阵修正结果都不能作为最终的 Q 矩阵，因为基于数据方法得到的 Q 矩阵可能与属性的认知结构不相符合，因此需要结合专家的意见。基于数据方法得到的 Q 矩阵可以为专家提供信息，为专家识别 Q 矩阵中的错误并进行修正提供支持。根据专家的意见，我们在讨论部分增加了相关的表述，详见文章讨论部分。

意见 2：由于文中仅研究了基于整体相对拟合统计量的 Q 矩阵修正方法，谨慎起见，建议作者换一个更恰当的题目。

回应：感谢专家的意见。参考专家的意见我们将论文题目修改为“基于类别水平的多级计分认知诊断 Q 矩阵修正：相对拟合统计量视角”。

意见 3：Ma 和 de la Torre(2019)研究中也使用了用 TIMSS 的 8 年级数学测试的数据，建议作者将本文研究结果与先前研究结果进行比较，这有助于读者理解本文的意义与价值。 建议：小修后再审

回应：非常感谢专家的意见，根据专家的意见，在实证研究部分我们补充分析了 Ma 和 de la Torre(2019)研究中使用的 TIMSS 的 8 年级数据，并与 Ma 和 de la Torre(2019)的研究结果进行了比较。研究发现在调整的属性上，BIC 方法调整的属性比 stepwise 方法稍多，两种方法修正后 Q 矩阵属性的一致率达 95%（结果详见文章表 7）；同时，在调整后 Q 矩阵的拟合上，stepwise 方法和 BIC 方法修正后的 Q 矩阵的绝对拟合指标和相对拟合指标均优于原始 Q 矩阵；而两种方法修正后的 Q 矩阵具有相近的拟合结果，在绝对拟合上，stepwise 方法调整后的 Q 矩阵 M_2 检验和 RMSEA 略优于 BIC 方法调整后的 Q 矩阵，但 SRMSR 指标以及相对拟合指标不如 BIC 方法调整后的 Q 矩阵（结果详见文章表 8）。

审稿人 2 意见：

意见 1：作者想把 Xu & Shang (2018)的基于二级计分的 Q 矩阵修正方法推广到基于 Seq-GDINA 模型的多级计分上，但存在不少问题。首先，原文对二级计分下 Q 矩阵修正的完备性和可识别性做了大量的推导证明，所使用的 EM 算法也给出了详细的推理过程，其方法很大程度上是基于比较复杂的正则化原理。这些关键的理论基础本文都没有，本文的理论部分比较薄弱和过于简单化。二级计分下成立的理论和算法并不必然在多级计分下可行，而且原文的饱和模型是 LCDM，与本文的 Seq-GDINA 存在较大差别，所以作者需要做大量类似的工作本文才具有理论上的说服力。

回应：非常感谢专家的意见。

首先，本文并不是将 Xu & Shang (2018)的方法推广到多级计分 CDMs 中，与 Xu & Shang (2018)相比，虽然本文方法也使用到了相对拟合统计量 BIC 指标，但两种方法是有区别的：（1）以某一题的 q 向量修正为例，Xu 和 Shang (2018)的方法需要通过估计的该题目参数（事先假定该项目测量了所有属性）来对该题的 q 向量进行推断，并将推断出的 q 向量作为候选矩阵与原有 q 向量进行比较，因此 Xu 和 Shang (2018)的方法是将由项目参数推导出的 q 向量与原有 q 向量进行 BIC 指标比较，并做出是否要修改 Q 阵的判断。而本文的方法不需要根据题目参数对 q 向量进行推断（也不需要事先假定项目测量了所有属性），在修正题目 j 第 h 类别时，是将所有可能的 q 向量分别作为题目 j 第 h 类别的 q 向量（其余题目 Q 矩阵不变），并根据 BIC 指标在所有可能的 q 向量中选择一个拟合最优即 BIC 最小的 q 向量。（2）Xu 和 Shang (2018)的方法使用 BIC 指标来决定是否更新 q 向量，每次修正只需要对题目参数推断出的 q 向量和原有的 q 向量进行比较。而本文中的方法在修正题目 j 第 h 类别时，需要对所有可能的 q 向量进行 BIC 统计量的比较。因此，在决定 q 向量是否需要更新时，前者只需要比较一个候选 q 向量和原有 q 向量的 BIC 指标大小，而本文的方法需要比较所有 q 向量与原有 q 向量的 BIC 指标大小。

总的来说，本文中的方法并不由题目参数来推断出 Q 矩阵，而是将所有可能的 q 向量分别作为题目 j 第 h 类别的 q 向量，在其他题目 Q 矩阵不变的情况下使用 BIC 指标为题目 j 第 h 类别挑选出能使模型相对拟合更好的 q 向量。因此两种方法的区别在于候选 q 向量的提出，前者是由题目参数推断出候选 q 向量，而本文是在所有可能 q 向量中使用相对拟合统计量选择最优的 q 向量。

其次，在对 Q 矩阵修正的研究中，我们发现其他同类研究（de la Torre & Chiu, 2016；Chen, 2017；Wang, Song, Ding, Meng, Cao, & Jie, 2018）未对 Q 矩阵的完备性和可识别性进

行数学推导，同样在 seq-GDINA 模型 Q 矩阵修正（Ma & de la Torre，2019）的研究中也未进行相关证明和推导。经与该文作者 Ma 联系和讨论，认为这是一个非常有意义也是非常复杂的问题，而对饱和模型（G-DINA）或多级计分模型（seq-GDINA）的 Q 矩阵完备性和可识别性未来研究可进行系统的数学证明推导，我们在文章最后的讨论部分也对此进行了讨论。此外，本文中模拟研究使用的 Q 矩阵和实证数据研究均来自于 Ma 和 de la Torre（Ma & de la Torre，2016；2019）的研究。

再次，本研究中 Q 矩阵修正的方法主要基于以下理论思考：当定义题目 j 的 q 向量时，在其他题目 Q 矩阵不变的情况下，在所有可能的 q 向量中能使模型相对拟合更好的 q 向量应该为题目 j 的 q 向量。在认知诊断中，常用的模型相对拟合指标包括-2LL、AIC 和 BIC。在约束模型（如 DINA）中，题目 q 向量标定错误会导致题目猜测参数和失误参数增加，从而降低模型的似然值，因此在约束模型中-2LL 指标可以挑选出恰当的 q 向量。而在复杂模型（R-RUM 或 GDINA）中，已有研究（Chen，de la Torre & Zhang，2013；Chen，2016）表明，若在原有 Q 矩阵基础上增加属性（overspecified）会产生更大的模型似然（由于模型参数个数增多）。所以在复杂模型中，-2LL 指标通常会挑选全为 **1** 的 q 向量($\boldsymbol{q}=[1111]$)作为题目 j 的 q 向量。因此在复杂模型中对题目 j 的 q 向量进行标定时需要对模型的参数个数进行惩罚，而 AIC 和 BIC 指标则是在-2LL 的基础上对模型的参数个数进行了惩罚。

本文方法实质上是使用模型相对拟合统计量，在只有题目 j 第 h 类别的 q 向量变化，其他条件不变的情况下进行的模型拟合度比较，由此为题目 j 第 h 类别选择出能使模型相对拟合最好的 q 向量。而在以往研究中，Chen，de la Torre 和 Zhang（2013）也将模型相对拟合指标（-2LL、AIC 和 BIC）用于对不同 Q 矩阵的比较，而 Xu 和 Shang (2018)的方法在对由项目参数推导出的候选 Q 矩阵与原有 Q 矩阵的比较时也是使用的相对拟合指标（BIC）。基于以上理论思考和以往研究，本文将相对统计量指标用于多级计分模型 Q 矩阵的修正。

因此本研究的出发点是在专家给定的 Q 矩阵基础上使用相对拟合指标来判断更新某些题目类别的 q 向量是否会使模型拟合更好，由此来进行 Q 矩阵修正。未来研究也可以考虑将 Xu 和 Shang (2018)的方法拓展到多级计分模型（seq-GDINA）中，并与本文中的方法进行比较。而对于多级计分模型中的 Q 矩阵的完备性和可识别性的相关推导及证明也有待研究者们的进一步研究。关于本文的方法与 Xu 和 Shang（2018）方法的区别我们也在文章的讨论部分进行了讨论，请见文章讨论红色字体部分，再次感谢专家的意见和建议。

意见 2：原文只推荐使用 BIC，本文加上-2LL 和 AIC 有何理论依据？需要给出相关的论证。还有原文的具体实施步骤与本文有较大差别，既不需要按特定的顺序去增加和删除属性，也没有穷尽和顺序算法的差别，作者同样需要给出理由。

回应：非常感谢专家的意见，本文受到 Xu 和 Shang（2018）以及 Chen，de la Torre 和 Zhang（2013）研究的启发，尝试将传统认知诊断中的模型相对拟合指标用于多级计分认知诊断模型 Q 矩阵的修正，其逻辑为当 Q 矩阵中题目 j 第 h 类别定义为不同的 q 向量，而其余题目的 Q 矩阵保持不变时，可以使用相对拟合指标来比较当题目 j 第 h 类别在不同 q 向量下的模型相对拟合度，从而选择相对拟合更好的 q 向量作为该类别的 q 向量。而-2LL、AIC 和 BIC 通常作为认知诊断中模型的相对拟合指标，因此这里考虑将三个相对拟合指标用于 Q 矩阵修正并进行比较。

如问题（1）所述，Xu 和 Shang（2018）由题目参数对 Q 矩阵进行推断，并使用 BIC 指标来判断候选 Q 矩阵与原有 Q 矩阵的优劣，以决定是否更新 Q 矩阵。而本文中的方法将所有可能的 q 向量分别作为题目 j 第 h 类别的 q 向量，在其他题目 Q 矩阵不变的情况下使用模型相对拟合指标挑选出题目 j 第 h 类别的 q 向量。理论上，题目 j 第 h 类别可以定义为任何潜在的 q 向量（2^K-1 种潜在 q 向量），直接的一种算法就是所有可能的 q 向量分别作为题目 j 第 h 类别的 q 向量，在其余题目 Q 矩阵不变的情况下分别进行模型估计，再计算相对拟合指标，这种方法比较耗时。而在实际中专家给定的 Q 矩阵中往往只包含少量的 Q 矩阵错误，因此顺序算法是考虑了专家提供的 Q 矩阵的信息。

顺序算法是先考察专家给定的 Q 矩阵中题目 j 第 h 类别的 q 向量是否有属性缺失，即分别将为 0 的属性变为 1 后是否能提高模型的相对拟合；再考察该类别的 q 向量中是否有属性冗余，即分别将为 1 的属性变为 0 是否能提高模型的相对拟合，由此来搜索合适的 q 向量。相对于穷尽算法，顺序算法不需要再对模型估计 2^K-1 次，大大减少了计算量和运行时间。

研究结果显示在复杂模型下，三种指标中-2LL 和 AIC 指标表现均不如 BIC 指标，而在约束模型（seq-DINA）中-2LL 与 BIC 指标是等价的，因此参考专家的意见以及表达的简洁性，我们将-2LL 和 AIC 指标的结果删除，只保留了 BIC 指标的结果。

意见 3：Ma & de la Torre (2019)已经发表过基于同样 Seq-GDINA 模型的多级计分 Q 矩阵修正方法。作者需要在理论上比较两种方法有何不同，尤其是提出新方法的必要性。同时需要通过模拟研究和真实数据比较两种方法在实证中的异同，尤其当结果相差较大时需要给出合

理的解释。

回应：Ma 和 de la Torre (2019)提出了 GDI 和 wald 检验相结合的 stepwise 方法进行 seq-GDINA 模型的 Q 矩阵修正，该方法先选择单属性 q 向量中具有最大 GDI 值的 q 向量作为基础，再考察属性的 wald 检验是否显著来决定是否增加或删除属性，并通过计算 wald 检验以后 q 向量的 GDI 指标来决定是否终止。该方法在确定每个类别的 q 向量时，需要进行多次 wald 检验，并计算标准误，计算相对复杂。而本文使用的 BIC 指标原理相对简单、计算也较简单，且在使用顺序算法（sequential search algorithm）以后大大减少了运算量和运算时间。此外，两种方法对 Q 矩阵修正的原理并不相同，Ma 和 de la Torre（2019）的方法中 wald 检验是从属性的角度来考察属性是否对答对概率有影响，GDI 指标则是保证 q 向量具有最大的区分度。而本文中的 BIC 方法是从整个模型拟合的角度来比较 q 向量增加或删除属性后模型的拟合，因此两种方法是从不同角度来考察 Q 矩阵是否合理。我们也在文章的讨论部分对两种方法进行了讨论，详见文章讨论部分。

根据专家的意见，我们在模拟研究和实证研究中增加了 Ma 和 de la Torre（2019）的方法，并与本文中的方法进行比较。研究发现在模拟研究中 BIC 方法的 PMR 和 AMR 指标在绝大多数实验条件下均优于 stepwise 方法；在修正后 Q 矩阵的绝对拟合上，BIC 方法修正后的 Q 矩阵的平均 RMSEA 指标也优于 stepwise 方法修正后的 Q 矩阵。实证研究发现 BIC 方法与 stepwise 方法修正后 Q 矩阵的绝对拟合指标和相对拟合指标均优于原始 Q 矩阵。

意见 4：研究所选取的真实数据仅有一题是多级计分的题目，并不能很好地说明该方法对于多级计分 Q 矩阵修正的效果，应尝试寻找更为恰当的真实数据，可参考 Ma & de la Torre (2019)的文章。

回应： 非常感谢专家的意见，本文采用的真实数据就是源自于 Ma 和 de la Torre(2019)的研究。根据专家的意见，我们又新增加了一个真实数据（Ma& de la Torre, 2016)分析，该数据包括 3 个多级计分题目，共 11 题，stepwise 方法和 BIC 方法分别调整了 17 个和 14 个属性，而 stepwise 方法调整后 Q 矩阵中属性 5（A5）没有被任何题目测量。在拟合上，BIC 指标修正后 Q 矩阵也比原有 Q 矩阵具有更高的绝对拟合指标和相对拟合指标。当然，未来研究还需采用更多的真实数据对本文中方法进行验证，我们也在文章的讨论部分进行了讨论。

第三轮

审稿人 1 意见：

意见 1：作者已经较好地回答了我提出的问题，同意推荐发表。

回应：非常感谢专家在前两轮提出的宝贵的修改意见，也感谢专家对修改后的文章的认同。

审稿人 2 意见：

意见 1：本文相对前文有了较大改进，不过与 Xu & Shang (2018)的联系仍没说清楚。虽然表面上本文对 q 向量的推断并不直接依赖题目参数，但其对 LL 的优化是建立在同时估算题目参数的基础上的，参考方程（5）和（6），这点本质上与 Xu & Shang (2018)一致。同时，Xu & Shang (2018)也有基于 provisional Q-matrix 进行逐题调整的方法，两文有不少相通之处，需要进一步阐明。两文最大区别并不是作者所说的两点，而是另外两点：1）二级与多级计分；2）一个使用结合 BIC 的、类 L1 的正则化方法进行优化调整，并具有理论证明，不需要估算所有可能的属性模式；另一个使用常规的优化方法估算所有的属性模式（我没理解错的话），然后用 BIC 挑选最好的。但如果 Xu & Shang (2018)的理论证明成立，我怀疑这两者基本一致，至少在二级计分上。

回应：非常感谢专家的意见和指点，根据专家的意见，我们在引言部分重新阐述了本文中的方法与 Xu 和 Shang (2018)方法的联系和区别。如专家所言，本文中的方法与 Xu 和 Shang (2018)方法有相似之处，都是需要对模型参数进行估计，并结合信息指标来进行 Q 矩阵修正。并且在修正 Q 矩阵时两种方法都是在其余题目 Q 矩阵保持不变的情况下，逐题（或类别）对 q 向量进行确定。而两种方法的区别在于，Xu 和 Shang（2018）的方法采用 TLP (truncated L1 penalty function)的正则化算法，由估计的项目参数稀疏矩阵来推断题目的 q 向量，并结合信息指标（BIC）来进行 Q 矩阵修正，因此并不需要对所有可能的 q 向量进行估计。而本文中的方法分别对所有可能的 q 向量进行估计，再通过信息指标挑选出最优的 q 向量。此外二者的区别还在于 Xu 和 Shang （2018）的方法是用于二级计分的 Q 矩阵估计或修正，而本文则是对多级计分 Q 矩阵修正进行研究。根据专家的意见，我们在引言部分进行了重新表述，再次感谢专家的意见和指点。

意见 2：顺序算法只考虑先增加再删除，建议同时比较相反和/或其他顺序，或者说明原因；

回应：非常感谢专家的意见，我们比较过四种顺序算法的表现，四种算法分别是本文使用的先增加属性再删除属性（forward-then-backward search algorithm, BIC_FB），和其他三种顺序

算法（1）先删除属性再增加属性（backward-then-forward search algorithm, BIC_BF）；（2）从单属性 q 向量中先挑选出 BIC 最小的 q 向量，然后再依次增加属性（forward search algorithm, BIC_F）；（3）从全为 $\boldsymbol{1}$ 的 q 向量依次删除属性（backward search algorithm, BIC_B）。其中前两种顺序算法是在专家给定 q 向量的基础上进行的搜索算法，而后两种算法则没有利用专家给定的 q 向量信息。

实验结果表明 BIC_FB 算法和 BIC_BF 算法的表现几乎一致，而 BIC_F 算法和 BIC_B 算法在一些实验条件下略低于 BIC_FB 算法，这也许是由于 BIC_F 算法和 BIC_B 算法没有利用专家给定的信息导致的。而其中 BIC_B 算法的表现最差，原因可能是 BIC_B 算法总是从全为 $\boldsymbol{1}$ 的 q 向量出发依次删除属性，而全为 $\boldsymbol{1}$ 的 q 向量由于参数个数更多可能会造成模型的估计误差变大。因此，考虑到四种算法的效果，以及文章的简洁性和文章的篇幅，我们在本文中仅报告 BIC_FB 算法的结果；当然，关于不同的顺序算法，我们也在文章的讨论部分进行了说明。再次感谢专家的意见。

意见 3：实验重复 100 次偏少，建议 200 次以上；FPR 好像与通常的定义反了；

回应：非常感谢专家的意见，根据专家的意见，我们进行了补充实验，将实验次数增加到了 200 次，且将文中的数据结果进行了更新。结果显示 200 次实验结果与 100 实验结果之间差异极小，并且没有改变原有实验结论。此外，我们也对文中 FRP（false-positive rate）进行了修改，谢谢专家的指正！

意见 4： 术语“测量模式”容易引起歧义，建议改为“属性模式”（我没理解错的话），或其他更明确的说法；

回应：非常感谢专家的意见，根据专家的意见，我们已经进行了修改。

意见 5： 与前人的联系最好在前面介绍清楚，不要留到后面讨论。

回应：谢谢专家的意见，我们已经将于前人方法的联系和区别在论文的引言部分进行了补充阐述，请见引言绿色字体部分。再次感谢专家的意见！