《心理学报》审稿意见与作者回应

题目：多级属性 Q 矩阵的验证与估计
作者：秦春影，喻晓锋

第一轮

审稿人 1 意见：

意见 1：多级 Q 矩阵的验证能够把学生诊断的分类从二分变为三级及以上的区分，诊断信息更加精细，因此具有重要的理论与实践意义。本文提出了两种多级 Q 矩阵验证的算法。在仔细阅读全稿后，需要作者从以下角度思考并回答我的一些疑问:

第一，衡量指标中的“迭代次数”是否具有重要意义。本文的两种算法是 EM，一般都 5 个迭代左右都能够收敛，所有的模拟和实证数据符合预期，但是左右优劣指标，具体几次收敛不是很有意义。也许运行时间更加重要。另外，这里的 OE 算法是一次增加一道新题，某种意义上来讲，OE 的迭代次数与 JE 的迭代次数含义不同（OE 是每个新题的收敛次数？），不具有可比性。

回应：非常好的意见。我们在修改的版本中，增加了运行时间这个指标，希望能够更清晰和全面地表达出本文方法在验证和估计 Q 矩阵时的表现。从运行时间来看， 虽然算法已经极大地降低了搜索的空间，但是该算法仍然是一个计算比较密集型的算法，各条件下的平均耗时仍然较高，未来需要考虑一些优化的技术，比如采用并行计算的方式，并将一些重复调用的模块改用执行效率更高的语言来编程。

对于 OE 和 JE 的迭代次数，它们确实有些不同，因为 JE 的迭代是在对所有的题目完成一次估计为一次迭代；OE 算法中的迭代是指完成所有的“新项目”估计后，如果“新项目”没有估计成功，则对包含“基础项目”和“新”项目的 Q 矩阵用 JE 算法进行联合估计，因此从这个角度来看，OE 算法中的迭代次数与 JE 算法中是一样的，也是指对所有项目完成一次估计的次数。它们的不同之处在于当 OE 算法对“新项目”成功估计时，则不需要用 JE 算法对包含“基础项目”和“新项目”的 Q 矩阵进行联合估计，这时的迭代次数就是 0。

我们对这部分内容进行了阐明，以便能更准确地表达。

意见 2：第二，OE 与 JE 设置条件的角度不同，因此可能存在等价性。比如，JE 中，在项目数为 15 时，是不是与”OE 中基础项目数为 10“时的若干情况存在等价（JE 中 15-5=10，相当于 OE 中基础为 10，但是需要同时处理 5 个新题的情况）。因此，可不可以把 JE 直接改造成等价形式的 OE，提高单题的估算准确性？ 或者把 OE 改造成等价情况下的 JE，同时处理多道新题（从 1 到提升到 2 题？3 题？4 题？ 还是 5 题？）。从这种观点来看，作者的模拟条件缺乏这些条件的通盘考虑，可能只是探索中了其中部分选项，有可能在模拟研究之前就认为排除了最高效的做法。希望作者从我这个角度，梳理有关的等价性，重新设计模拟研究。

基于以上意见，建议大修后再审。

回应：对于 OE 算法和 JE 算法，我们设置的出发点是两种实际 可能存在的测验情形，其中 JE 对应学科专家已经对测验的 Q 矩阵进行了界定或不同专家间的意见不一致的情形，我们需要对其进行验证或从多个不同的备选 Q 矩阵中做出判断，因此，这种情况下，对应的是“初

始 Q 矩阵”中存在部分错误的情形；而对于 OE 算法，我们设置的出发点是已经有少部分题目的属性设定得到了确定，有一批新的题目需要进行标定或入库的情形，非常类似于很多文献中提到的“在线标定”。
关于二者的不同之处，我们借鉴您举出的例子进行说明。对于 JE 算法，当项目数为 15，错误标定项目数为 5 时，它与 OE 算法中基础题为 10，新题为 5 的情形还是有不同的，理由是：JE 算法是将 15 道题进行联合估计，估计的时候并不知道哪道题是标定错误的，因此会对所有的 15 道题进行估计；而 OE 算法是固定 10 道基础题，按照增量的方式每次逐步加入一道新题进行估计，只估计加入的这一道新题。这也是 OE 算法结果看起来更好的原因，当 OE 算法一次性估计多道“新题“的做法，我们是做过实验的，效果并没有一次加入一道新题的效果好，因为那样会带入更多的噪音而产生更大的”遮罩效应“(Fung, 1993; Yuan & Zhong, 2008)。对于如何保证“基础题“的界定是正确的，则又是需要进一步研究的问题。

因此，这两个算法对应了两种实际可能的应用场景，我们在考虑后还是把它们分别按两个研究进行设计。以上是我们的理解和回复，如果不能回复您的顾虑，请指出。再次感谢您的建议！

审稿人 2 意见：

意见 1：论文将基于 S 统计量的 Q 矩阵估计方法拓展到了多级属性下，有一定的理论创新和对往后研究的指导意义。论文的总体撰写上比较清晰，使用方法正确，结论较为可靠。但评审人仍有一些论文的完善建议：
定理 1 无须证明，给出结论就可以了。
回应：我们在修改版中删除了定理 1，改为给出相应的结论并做必要的阐明。

意见 2：文章将基于 S 统计量的方法进行了多级推广，那么这个指标的提出部分应该就是研究的重要部分，也就是当前论文的 4 部分。但无论是前面的 1 部分还是 2 部分、3 部分都是对前人研究的回溯。4 部分不适合与其并列。建议将 2、3 部分都并入到引言 1 部分，4 部分与引言并列，重点阐述提出的指标。
回应：非常好的意见！我们对于前面 4 部分的结构进行了较大的调整，将之前的第 2 部分与第 3 部分进行了重写，并合并成为新的第 2 部分，将最重要的指标的提出作为第 3 部分，重点对所提出的指标进行详细的阐明，进行了更为完整的介绍，对它的性质也进行了更进一步的分析和讨论。以期能够更为全面和清晰地描述指标及其使用场景。

意见 3：本文的 JE 和 OE 算法都是基于 S 统计量的，引言部分也提到了为什么选择基于 S 统计量的方法，但在讨论部分建议增加一部分基于其他方法的讨论
回应：我们在讨论部分补充了对其它方法或其它潜在的方法的讨论，补充了一些比较新的研究文献。

意见 4：在 5.1 中项目个数两个水平分别为 15 和 30，错误的项目个数水平分别为 3，4，5。取这些水平的依据是什么？
回应：我们补充说明了上面提到的实验条件设置的考虑或依据。一方面本研究采用的多值属性诊断模型来自 Chen 和 de la Torre(2013)；对于 Q 矩阵估计，拓广的是 Liu, Xu 和 Ying(2012)中的方法；另一方面，目前还没有关于多值属性 Q 矩阵估计有关的文献公开出版。因此本

文研究中所使用的项目个数和错误的项目个数的设计，对于 JE 算法，我们参考的是 Liu 等人(2012)和 Qin 等人(2020)的设计，两种测验长度的设计我们参考的是本文模型对应的提出文献（Chen & de la Torre, 2013），分别是 30 和 15，3 种错误项目个数，分别是 3，4 和 5；而对于 OE 算法中的错误项目个数，我们参考的是 Qin 等人(2015, 2020)的设计，考虑了基础项目个数分别为 8，9，10，11，12，13，14，15 共 8 种情形。对于以上的内容，我们补充了相应的说明。

意见 5：分别考虑了两种错误类型单独存在时的情况，在模拟研究中分开探讨有利于梳理出清晰的影响结果。而实际情况中，往往可能两种错误类型同时存在，如果两者同时出现是什么情况？又或者这是否是实际使用不佳的原因，可适当进行讨论。
回应：错误类型 I 是简单的情况，即只存在属性值在大小上标定错误，即过高（但不包括本身是 0 的属性设定为其它值）或过低（但不能低至 0）的设定；错误类型 II 是复杂的情形，即不但包括错误类型 I 的情况，而且包括将本身是 0 的情形设定为非 0，也包括将非 0 的属性值设置为 0。因此，错误类型 II 是实际测验中更可能存在的情形，属于更一般的错误类型，而错误类型 I 是错误类型 II 的特殊情况。
我们对于这一点进行了更具体的阐明。

意见 6：论文图形的区分不太清晰，可尝试或者提高分辨率或者尝试更换图例。
回应：我们对论文中的图形和图例都进行了调整，提高了图形的质量。

意见 7：文中存在部分笔误，请仔细检查核对。
回应：我们对全文进行了仔细的核查，对存在的部分笔误进行了修改，对一些内容进行改写。具体请见正文中高亮的部分。

第二轮

审稿人 1 意见：
已无其他修改意见，同意发表。

审稿人 2 意见：
无进一步意见

编委意见：
这篇论文将基于 S 统计量的 Q 矩阵估计方法拓展到了多级属性下，有一定的理论创新，论文的总体撰写上比较清晰，结论可靠。经审稿人评审，作者修改，论文达到发表要求。推荐发表。

主编意见：
该论文经两外审专家评审，一审均提出一些需要澄清和解释的问题，作者对这些问题进行了详尽的回答并相应对论文进行了修改，审稿人表示认可，在第二轮审稿中均同意发表。整个审稿流程规范，同意发表。