《心理科学进展》审稿意见与作者回应

题目：认知诊断评估中 Q 矩阵理论及应用
作者：宋丽红 汪文义 丁树良

第一轮

审稿人 1 意见：

本文详细回顾了认知诊断测评中有关 Q 矩阵完备性和模型可识别性等相关内容，非常有益于领域新人全面系统学习 Q 矩阵理论。整体而言，文章内容翔实，撰写逻辑清晰。但“综”的内容多，“述”的内容少，建议适当平衡两者的篇幅：

意见 1：综述动机可进一步丰富，比如，Q 矩阵不完备，或者在模型不可识别情况下，会带来什么危害？是否有研究犯过这类错误而导致了具体的危害？如果没有危害，似乎没有探讨他的必要性；

回应：感谢您的重要意见。完备 Q 矩阵主要用于指导认知诊断测验设计，并且已有研究证实 Q 矩阵不完备会引起严重的分类问题并降低诊断分类结果的准确率，Q 矩阵设计还会影响模型可识别性和导致部分参数无法得到唯一估计结果。我们在引言中增加了下面两段，重点叙述了 Q 矩阵设计的重要性和不完备 Q 矩阵会带来的危害，以及模型不可识别性的后果：

Q 矩阵设计是认知诊断测验设计的十分重要方面（丁树良等人，2011；Liu et al., 2016；Madison & Bradshaw, 2015; Tian et al., 2020; Tu et al., 2019）。诊断测验设计核心任务之一设计测验各个题目所测量的哪个（些）属性，也就是解决 Q 矩阵设计或测验蓝图问题（Leighton et al., 2004），Tatsuoka(2009)提出充分 Q 矩阵用于指导认知诊断测验编制。完备 Q 矩阵作为一种重要 Q 矩阵设计，对于提高分类准确率具有重要作用。DeCarlo（2010）在分析分数减法数据时发现，不完备 Q 矩阵会引起严重的分类问题，测验 Q 矩阵设计不当，测验为被试在某些属性上带来的信息甚至还不如先验信息。丁树良等人（2011）研究发现，完备 Q 矩阵（至少含一个可达阵）比不完备 Q 矩阵的模式判准率高出 20%以上；Tian 等人(2020)发现完备 Q 矩阵可提高纵向诊断分类模型的分类准确率。Madison 和 Bradshaw（2015）比较了不同 Q 矩阵设计对分类准确率的影响，相比其他不完备 Q 矩阵，包含每个属性单独测量 1 次或 2 次的 Q 矩阵（完备 Q 矩阵）在参数估计算法收敛性、属性分类准确率和属性信度方面均具有明显优势。Kuo 等人（2016）例子显示，在线性属性层级结构下，基于认知诊断指标或属性诊断指标选择试题，所得测验 Q 矩阵不完备，由此他们提出了具有更高判准率的

组卷方法。

Q 矩阵设计还与认知诊断模型识别问题和参数估计量的一致性密切相关。统计模型可识别，是得到参数一致估计和有效推断结果的必要条件，也是获得可靠且有效评价的基础(Gu & Xu，2019a)。Q 矩阵不完备会引起知识状态等价类，即造成同一等价类中多个知识状态的概率参数不可识别，还会导致 Q 矩阵估计不可识别。

意见 2：介绍每种方法前，是否可以简单阐述下原文作者的研究动机，以便读者理解为什么要去学习该方法。

回应：感谢您的建设性意见。在每种方法前，增加了简要叙述原文作者的发现的问题或研究动机的内容，便于读者阅读和建立各种方法之间的联系，下面举例说明增加的部分内容。在介绍多值充分 Q 矩阵之前，增加了下面内容说明原作者的研究动机：不同项目对同一属性的认知水平要求不尽相同，如果二值 Q 矩阵不能很好地反映项目中同一属性的难度水平或认知水平高低的差异，并且复杂问题解决过程中属性粒度定义不能太细（粒度过细会导致属性数过大，并且属性数多时层级结构也更难确定），这时测验 Q 矩阵就需要采用多值 Q 矩阵。采用专家所定义的多值属性（Chen & de la Torre, 2013），可以指导测验设计和事先针对性设计和开发测量特定属性水平的项目。受二值完备 Q 矩阵的启发，Sun 等人（2013）在开发多级评分广义距离判别法（GDD-P）时率先提出了多值充分 Q 矩阵。

在介绍知识空间理论下完备 Q 矩阵之前，增加了以下内容说明原作者的意图：Heller(2022)对知识空间理论下的完备 Q 矩阵的相关结论进行总结与梳理。与前面讨论的独立结构和属性层级结构不同，知识空间理论下的完备 Q 矩阵适合于更一般属性结构。独立结构和属性层级结构下知识状态全集对交和并运算封闭，而一般属性结构下只要求对并运算封闭。因为属性结构发生了变化，完备 Q 矩阵的条件也有所不同。

在介绍部分识别条件之前，介绍了严格识别条件过强的研究动机：因为许多真实测验的 Q 矩阵并不满足如此强的严格可识别条件，Gu 和 Xu(2020)、Gu 和 Xu(2021)研究 部分识别（partial identifiability），尝试给出 Q 矩阵更为实用的约束条件。

在介绍属性层级结构下模型可识别条件之前，介绍了原文作者考虑的研究问题和研究意图：上面叙述的模型严格识别、部分识别、一般识别的结论都是在属性相互独立条件下所给出的，且属性层级结构下完备 Q 矩阵仍不能保证 可识别或者直接从数据中估计，于是 Gu 和 Xu (2021b, 2023)探讨属性层级结构下 Q 矩阵满足什么条件时模型可识别问题。

意见 3：特定方法中的次要的公式或例子数量可以考虑放到附录部分。

回应：将完备 Q 矩阵的检查过程列入附录 1，以及将模型识别概念和条件列入附录 2，附录

2 中列入了较为复杂的 6 个定理及其相应例子。同时重点修改了正文中相关内容，保证上下文内容的衔接性。

……

审稿人 2 意见：

Q 矩阵理论是认知诊断评估的核心成分之一，本文详细地叙述了近年来 Q 矩阵理论及其应用，重点介绍了 Q 矩阵设计的重要结论，并结合具体的例子介绍如何使用，可供认知诊断理论研究者和实践者参考，仍有以下建议供作者参考：

意见 1：除了通过具体的例子说明重要结论如何使用之外，建议在研究结论部分用表格呈现 Q 矩阵设计或选择的使用条件、情景、具体方法、和 CDM 结合需要注意的问题（什么情景下该使用怎么使用 CDM 或进行诊断），以及基于此能得到的主要结论，这样实践者才能按图索骥。

回应：感谢您的建设性意见。在结论部分增加了表 2 和一段描述。表 2 详细地列出了 Q 矩阵特点、满足条件、应用情景和推荐的诊断方法。不同定义下的完备 Q 矩阵分别可用于测验不同阶段并有着不同作用，以及伴随着推荐的认知诊断模型或方法。基于理想反应模式的完备 Q 矩阵首先可用于指导测验设计，在给定属性及其层级关系以后，可以根据属性层级结构设计完备 Q 矩阵，并用于指导测验题目编制。在收集到实测数据之后，在小样本量情景下，可采用非参数认知诊断方法；在样本量中等情景下，可采用 DINA 或 DINO 模型；在大样本量情景下，可以选用一般化认知诊断模型（GDINA、LCDM、GDM），借助模型可识别条件并结合数据分析，判断 Q 矩阵、项目参数、分布参数、属性结构等参数的可识别性。

意见 2：在引言、讨论中需要重点强调 Q 矩阵在测验设计中的重要指导作用。在讨论部分，涉及到开发新模型时，研究者也应关注 Q 矩阵设计。

回应：在引言中增加了下面一段，重点强调了 Q 矩阵在测验设计中的重要指导作用，以及不完备 Q 矩阵会带来的危害：

Q 矩阵设计是认知诊断测验设计的十分重要方面（丁树良等人，2011；Liu et al., 2016；Madison & Bradshaw, 2015; Tian et al., 2020; Tu et al., 2019）。诊断测验设计核心任务之一设计测验各个题目所测量的哪个（些）属性，也就是解决 Q 矩阵设计或测验蓝图问题（Leighton et al., 2004），Tatsuoka(2009)提出充分 Q 矩阵用于指导认知诊断测验编制。完备 Q 矩阵作为一种重要 Q 矩阵设计，对于提高分类准确率具有重要作用。DeCarlo（2010）在分析分数减

法数据时发现，不完备 Q 矩阵会引起严重的分类问题，测验 Q 矩阵设计不当，测验为被试在某些属性上带来的信息甚至还不如先验信息。丁树良等人（2011）研究发现，完备 Q 矩阵（至少含一个可达阵）比不完备 Q 矩阵的模式判准率高出 20%以上；Tian 等人(2020)发现完备 Q 矩阵可提高纵向诊断分类模型的分类准确率。Madison 和 Bradshaw（2015）比较了不同 Q 矩阵设计对分类准确率的影响，相比其他不完备 Q 矩阵，包含每个属性单独测量 1 次或 2 次的 Q 矩阵（完备 Q 矩阵）在参数估计算法收敛性、属性分类准确率和属性信度方面均具有明显优势。Kuo 等人（2016）例子显示，在线性属性层级结构下，基于认知诊断指标或属性诊断指标选择试题，所得测验 Q 矩阵不完备，由此他们提出了具有更高判准率的组卷方法。

在结论、展望中也多处强调了完备 Q 矩阵在测验设计中的重要指导作用，并在展望中第 2 段中增加了叙述“另外，在新开发认知诊断模型时，也要注意 Q 矩阵设计，以保证新模型各类参数可识别。”

意见 3：文中提到定理 13 相对较难判断，而基于 Q 矩阵的定理 14 较容易使用，建议省略定理 13。此外，为了考虑到大部分读者的接受性，对于一些复杂的证明建议挪到附录中呈现，正文主体部分建议作者能够用一个例子贯穿始终，从开始诊断，到诊断过程，到最后需要注意的问题都有哪些。相当于做一个 Didactic，提供模板化的内容。

回应：为保证内容的联贯性，首先将完备 Q 矩阵的检查过程列入附录 1。根据您的建议，将稍复杂的模型识别概念和条件列入附录 2，附录 2 中列入了较为复杂的 6 个定理及其相应例子。同时重点修改了正文中相关内容，以保证上下文内容的衔接性。另外，对于前提条件相同的结论，基本上使用了相同 Q 矩阵例子。例如，在 2.3 节中使用了一个延续的例子，在 3.2 节的两个小节中使用了两个延续的例子。

意见 4：对于定理 17，建议也补充例子进行说明。

回应：感谢您的十分具体的意见。增加了原定理 17 应用的例子，同时定理 17 修改为附录 2 中定理 A6，并在其后增加了应用例子，例子请见附录最末。

意见 5：“一般可识别性（generically identifiable）”的英文无需重复出现。

回应：感谢您的认真细致的建议。出现冗余的“一般可识别性（generically identifiable）”的英文已经删除，同时仔细检查和修改了全文相关描述、记号、公式、简写等内容。

第二轮

审稿人 1 意见：

感谢作者对本人意见的回复和修改。最后再提一个小建议，在正文开头部分，应该进一步指出认知诊断的实用价值。作者仅用了一句话阐述认知诊断测评广泛应用于一些领域，但用的是否有效，是否有价值等并没有给予说明，建议补充。

回应：感谢您的意见。根据您的建议，在正文第一段第一句之后增加了以下补充说明：众多研究显示(王立君 等, 2020; Toprak, 2021; von Davier & Lee, 2019)，认知诊断在学习系统中学习者弱项诊断、报告反馈与资源推荐，在大规模评价数据分析与细粒度诊断，在识别问题解决策略和职业教育，在教学干预方法或个性化补救教学效果评价等方面都发挥着重要作用。

审稿人 2 意见：

作者较好的回答了审稿人的问题，增强了文章的可读性，论文质量有了较大提升。仍然有两个小问题和作者讨论：

意见 1：表 2 中好像缺少“独立-多分属性”情景下的应用条件，并且属性的认知机制间的关系划分应该是：补偿-非补偿，链接-非链接的关系，似乎在表 2 中没有穷尽，是有些条件下不能满足导致吗？此外，除了饱和模型和两个典型的 DINA 和 DINO 模型外，还有 ACDM/LLM 等加法模型，该如何使用呢？

回应：谢谢您细心评阅。根据您的意见，因为“独立-多分属性”可视为“层级-多分属性”的特例，故将此条件补充到表 2 中。根据认知机制的分类(von Davier & Lee, 2019)，因为连接或非补偿(conjunctive or non-compensatory)、非连接或补偿(disjunctive or compensatory)经常互用，故原表 2 中用了连接和补偿。根据您的建议，为了统一用两个相互对立的连接和非连接概念，故将表 2 中“补偿”修改为“非连接”。已有文献尚未对所有组合条件进行研究，比如 Gu 和 Xu (2021b, 2023)尚未研究属性层级结构下一般化认知诊断模型可识别的条件，以及 Heller(2022)仅给出一般结构下非连接机制的结论，故表中并没有穷尽所有组合条件。另外，在表 2 中，针对非连接机制补充了完备 Q 矩阵的两个充分条件，以及加性认知诊断模型(the additive cognitive diagnosis model, ACDM)或线性逻辑斯蒂克模型(linear logistic test model, LLTM)使用的条件。

意见 2：在现实应用中，CDM 很难满足大样本情景，甚至中等样本可能都无法满足，根据表 2 梳理可知，小样本作者均推荐的是 NPC，但有很多研究在 500 人以内也在使用如 DINA，GDINA 等模型进行分析，是否可以认为这些研究都用错了模型呢？而应该使用 NPC 才对。不过 Ma 等（2021）提出了基本小样本的贝叶斯估计，其研究可以处理 30 人的小样本。表 2 中是否涵盖了这种情况呢？【Ma, W., & Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples Bayes modal estimation and monotonic constraints. Applied Psychological Measurement, 45(2), 95-111.】

回应：这是一个非常有趣的问题，值得深入研究。在小样本量情形下，可能使用 DINA，GDINA 等模型进行分析，仍可以获得一些有价值的结论。在表 2 中，小样本组合条件下均推荐 NPC，这主要有三方面考虑：第一，因为此时所伴随列出的测验 Q 矩阵要求较低（仅要求包含单位阵或可达阵），这尚不能满足 DINA 或 DINO 模型参数严格或部分可识别的条件；第二，样本量 500 基本上是认知诊断模型获得较高精度时对样本量的最低要求，这是众多研究形成的共识(参见：Sen & Cohen, 2021)。虽然融入先验分布信息的贝叶斯估计方法可以加速算法收敛，但是样本量 500 比样本量 30 或 100 的模式判准率至少高 20%或 10%(Ma & Jiang, 2021)；第三，根据最新研究（Ma et.al., 2023），NPC 和拓广的 NPC（the general NPC, GPNC）仍是小样本量情形下推荐方法。

第三轮

审稿人 2 意见：同意发表。

编委 1 意见：

意见 1：引言部分可以简单一点，删除可有可无的内容。

回应：根据您的建议，将引言中原第 2 段内容进行了简化，精简成两句分别放到引言中第 1 段后面和新第 2 段前面，起到承上启下的作用。还对引言中内容联系密切的原第 4 段和第 5 段内容进行了简化，并合并 1 段。同时，精简了其他内容的叙述。

意见 2：“知识空间理论研究团队也一直研究完备 Q 矩阵(Heller, 2022)，鉴于认知诊断中完备 Q 矩阵在结构表征、测验设计、模型识别、诊断分类等方面的重要作用，并且诸多研究团队不断深入研究 Q 矩阵完备问题。”句子不通顺，请重新表述。

回应：谢谢您细致审阅。该句修改为：“Q 矩阵在结构表征、测验设计、模型识别、诊断分类等方面具有重要作用，并且诸多研究者长期深入研究 Q 矩阵理论并取得了大量成果，但目前缺乏相关的文献综述与评论。本文重点梳理近 15 年 Q 矩阵理论和应用”。另外，将“知识空间理论研究团队也一直研究完备 Q 矩阵(Heller, 2022)”列入引言中第三段相关内容中。

意见 3：本文象教科书一样将定义和定理单列成段，或许可以引起读者注意，但要与上下文自然衔接（请参考此处修改建议），也方便读者看到定义或定理是谁做出的。

回应：已经按照您的批注进行了 2.1.1 节的内容修改。同时，为了保证全文中单列成段的定义或定理与上下文自然衔接，其他相关部分，如 2.1.2 节、2.3 节、2.4.4 节，也遵照您的这一思路进行了修改，更好地引用定义和定理出处，并且使得内容之间的逻辑联系更为紧密。

意见 4：这一小节讲充分 Q 矩阵，下面一节才讲充要 Q 矩阵。，

回应：根据您的意见，该句修改为：借助充分 Q 矩阵要满足的条件指导测验设计和项目开发，使得测验真正测到所要测量的结构和属性，从而提高测验的结构效度。

编委 2 意见：同意发表。

第四轮

主编意见：根据编委和审稿专家的意见，建议发表。