《心理学报》审稿意见与作者回应

题目：基于可达阵的多级评分最简完备 Q 矩阵的设计
作者：唐小娟；彭志霞；秦珊珊；丁树良；毛萌萌；李瑜

第一轮

审稿人 1 意见：

Q 矩阵设计的好坏会影响诊断分类的结果，针对当前少有多级计分情境下的 Q 矩阵设计，论文《基于可达阵的多级评分最简完备 Q 矩阵的设计》基于 RP-DINA 模型，提出了多级计分的最简完备 Q 矩阵的设计方案，研究具有一定的价值。但研究还存在以下值得进一步探讨的地方：

回应：非常感谢专家对本研究的肯定，我们将认真学习领会专家的意见与建议，并按照专家的指导意见对全文进行修改，努力将文章质量提升到专家期望的水平。

意见 1： 文献综述不全，当前除了丁树良等的研究外，还包括 Culpepper(2019)、Fang 等(2019)，Y.Chen 等(2020)、Liu 和 Xu(2023)以及 Ouyang 和 Xu(2022)均讨论了多级计分情境下的 Q 矩阵可识别问题。除非作者认为 Q 矩阵可识别问题和 Q 矩阵完备性问题是不同的问题，否则该文章的文献综述是不全的。

Chen, Y., Culpepper, S., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis. Psychometrika, 85 (1), 121–153.

Culpepper, S. A. (2019). An exploratory diagnostic model for ordinal responses with binary attributes: Identifiability and estimation. Psychometrika, 84(4), 921-940.

Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. Psychometrika, 84 (1), 19–40.

Lin, M., & Xu, G. (2023). Identifiability of Cognitive Diagnosis Models with Polytomous Responses. preprint arXiv:2304.01363.

Ouyang, J., & Xu, G. (2022). Identifiability of latent class models with covariates. psychometrika, 87(4), 1343-1360.

回应：非常感谢！专家提供的这些参考文献，为我们提供了更宽广的视野。我们根据专家提供的文献，对可识别性概念及其相关研究重新进行了梳理，并回答了 Q 矩阵可识别性与 Q 矩阵完备性之间的关系（见引言部分蓝色字体）。具体内容如下：

（1）在认知诊断测验中，刻画项目与属性关系的 Q 矩阵对被试分类的精度有着非常重要的影响。通过 Q 矩阵，诱导出被试的差异性作答，差异性越大越有利于认知诊断模型精准地识别被试，故基于 Q 矩阵的认知诊断模型可识别问题受到广泛关注。可识别研究内容主要包括 Q 矩阵完备性、基于 Q 矩阵的模型参数估计以及 Q 矩阵和模型参数的联合估计。在已有的研究中，可识别的充分或（和）必要条件之一是包括 Q 矩阵为完备矩阵，也即 Q 矩阵的完备性是认知诊断模型具有可识别性的关键条件。因此，Q 矩阵完备性问题的研究是 Q 矩阵可识别问题的一个子问题。

（2）到目前为止，关于多级评分完备 Q 矩阵测验设计的研究很少。本文关于完备 Q 矩阵的讨论，主要考察 Q 矩阵否能够建立理想反应模式（IRP）与知识状态的一一对应关系。囿于我们的视野，基于理想反应模式的完备 Q 矩阵的研究，仅发现丁树良、罗芬等人（2014）和

丁树良、汪文义等人（2014）多级评分完备 Q 矩阵设计及其理论证明。
（3）根据已有研究，可识别的条件之一为完备 Q 矩阵必含单位阵。但含单位阵使得测验比较单一，且当考虑属性层级结构时，可识别性条件并不十分明确（Gu 和 Xu，2021）。本文提出的最简完备 Q 矩阵所需项目比单位阵少，可以减少计算的复杂度，降低项目曝光率，增加了测验设计的多样性。

意见 2：目前，关于 Q 矩阵能否完全识别属性掌握模式的相关研究大体可归纳成两类：一个类以 Chiu 等人为主，他们将该问题定义为 Q 矩阵是否完备；而以 Xu 等则将该问题定义为 Q 矩阵的识别性问题。这二者之间是否存在本质差异？若存在本质差异，作者为何关注的是完备性，而不是可识别性？据审稿人所知，当前 CDM 领域的研究者，在应用 CDM 处理数据时，多关注 Q 矩阵的可识别问题。
回应：根据您的提出的问题，我们将相关的知识进行了整理。如前所述，我们认为，Q 矩阵的完备性问题是 Q 矩阵可识别性中的一个子问题。这一理解可能不够深入，若有不妥，还请专家指正。

意见 3：引言部分（第 4 页），作者指出“… 不能解决复杂结构。的测验设计问题…”，什么样的结构可以称为复杂结构的测验设计？作者在本文当中是否解决了复杂结构的测验设计问题？
回应：丁树良、罗芬等人（2014）和丁树良、汪文义等人（2014）只给出几种基本属性层级结构（线型、收敛型、分支型、无结构和独立结构）相对应的完备 Q 矩阵。在实际应用中，属性层级结构更为复杂，通常由这些基本属性层级结构复合而成，而如何得到这些复杂结构的完备 Q 矩阵，还未有相关研究（见引言倒数第二段蓝色字体）。本文提出了计算最简完备 Q 矩阵的通用算法，可适用于包括复杂结构在内的各种属性层级结构，如 Kohn 和 Chiu(2021) 给出 11 个属性的复杂的属性层级结构（见第 2,3 部分）。

意见 4：DINA 模型是限制性非常强的 CDM，其适用范围非常受限，因此其普适性亦不强。若作者以该模型为主要模型，则该研究的研究意义将极大受限。
回应：感谢专家的提问。含单位阵的完备 Q 矩阵均能建构理想反应模式与知识状态的一一对应关系，这是在没有参数影响下的基于 Q 矩阵的可识别问题；反之，如果在这种情况下知识状态不能被识别，那么加上参数影响，则知识状态更难以被识别。故本文关于完备 Q 矩阵的讨论，着重考察该矩阵是否能够建立理想反应模式（IRP）与知识状态的一一对应关系，故此阶段完备性的讨论与认知诊断模型无关（当然 IRP 的建立与测验 Q 矩阵和评分规则有关，有的评分规则由认知诊断模型规定）。进一步地，在文章中，我们结合认知诊断模型对 Q 矩阵的完备性进行了验证，发现基于 GPDINA 模型，只要主效应存在，由本文例题中得到的最简完备 Q 矩阵仍是完备 Q 矩阵，这一点我们在文章第 4 部分进行了修改（见蓝色字体）。此外，从现有文献来看（包括专家所提供的文献），其中的一些研究也是基于 DINA 或 GPDINA 模型的可识别性研究，而 GPDINA 模型可转化为多种认知诊断模型。未来，我们将按照专家的提示，进一步研究基于其他认知诊断模型的完备性问题。

意见 5：本文使用的计分方式是题目考察几个属性便得几分的规则，这种规则在模拟条件下容易实现，但在实际情境中，这种记分却不常见，使用这种不常见的计分规则，其实践意义该如何保证？
回应：感谢专家的提问，本研究选取做对一个属性便多得一分的规则主要有两点原因：
（1）因为这是最早提出的多级评分规则(Tatsuoka 1985,1995)，也是在理论研究中运用较多

的评分规则(Tatsuoka 1985,1995；丁树良，罗芬等, 2014; 丁树良，汪文义等, 2014；康春花等，2013; 田伟，辛涛，2012；祝玉芳，2009)。文中提到，一般地，根据不同的评分规则，提取的完备 Q 矩阵是不同的。对于多级评分来说，可达阵也是基于这种评分规则才具有完备性，否则可达阵也不一定是完备 Q 矩阵。在实际情境中的评分规则没有有一个通用的表示方法，方法众多，我们难以针对每种评分方法给出具体的完备 Q 阵，故只能针对较为经典的评分规则，进行分析。研究者只要能够按照文中完备性定义，建构理想反应模式和知识状态的一一对应，则同样可以提炼出完备 Q 阵。
（2）在研究过程中，我们还发现，当每个属性权重一样且每题满分值大于该项目考察属性总数时，由本文得到的完备 Q 矩阵仍然为完备 Q 矩阵，这就扩大了本文方法的适用范围。（修改内容见 6.1 蓝色字体）

意见 6：作者所研究的 Q 矩阵是题目水平的，还是类别水平的？即每个评分类别对应一个 q 向量（类别水平），还是一个题目对应一个 q 向量（题目水平）？
回应：本文使用的是项目水平，也就是一个题目对应一个 q 向量。 按照专家的提示，我们在正文中增加了此内容（修改内容见引言最后一句蓝色字体）。

意见 7：最简完备 Q 矩阵是 Q 矩阵可识别的充分条件？还是必要条件？亦或是充分条件？
回应：本文提出的最简完备 Q 矩阵概念是建立理想反应模式与知识状态一一对应关系，是可识别知识状态的充分条件。

意见 8：根据研究一的结果，要想获得较高的分类准确性，需要保证题目均有较高的质量，但这在实践情境中几乎是不可能的任务，如此的话，本文研究的意义又在哪里呢？
回应：感谢专家的指正。本研究的初衷是希望用较少的题对被试做最大的区分，这种设计尤其适用于随堂测验之类的情境，需要用较少且优质的题目一一区分被试，故而采用质量较高的题目。当然，含最简完备 Q 矩阵的测验具备完备性，也可以运用于题量较大的测验。作为一项科学研究，我们通常从最理想的状况出发，然后再考察在不同条件下原有理论的鲁棒性。根据专家意见，我们设计了适合终结性评估和过程性评估的两种测验，即设计题量为 40 题且包含一个或多个最简完备 Q 矩阵和可达阵的测验，以及题量等于可达阵列数的且只有一个或多个最简完备 Q 矩阵和可达阵的测验。同时，修改了参数，长测验选取了高、中和低质量的题目进行模拟，短测验选取了质量高和中等的题目以适应随堂测验等小测验的需求，研究结果表明，原有结论依然有效（见第 5 部分蓝色字体）。

意见 9：文章缺少实证研究，因而无法知晓作者所提方法在实践情境中的效用。
回应：感谢专家的建议。由于短时间内采集实测数据比较困难，我们联系国内同行专家，获得同行专家提供的实证数据支持，通过对这些实测数据的分析，结果表明，结构化最简完备 Q 矩阵与可达阵的识别重复率达 90%以上，且最简完备 Q 矩阵识别的知识状态类型更多一些（见第 6 部分蓝色字体）。

意见 10：讨论与展望部分建议给出更加直观的建议，如在实际测验中如何进行 Q 矩阵设计以平衡时间和经济成本？等等
回应：非常感谢专家的意见和建议。我们进一步丰富了讨论部分有关 Q 矩阵设计的建议，针对在研究中出现的问题进行了详细说明：比如增加了满分值对完备 Q 矩阵的影响；修改了测验的投入与产出比（计算判准率与时间和题量的比）计算公式；修改了非结构化完备 Q 阵说明等。

……

审稿人 2 意见：

本研究提出了多级计分中最简完备 Q 矩阵的设计方法，为认知诊断题库建设以及短题长的自适应考试选题提供了思路。具体来说，本文提出了，结构化与非结构化最简完备 Q 矩阵的两种设计方法，并使用模拟研究验证了其有效性。本人有以下问题供作者思考：

意见 1：模拟研究的项目参数设计过于理想化。现实的认知诊断参数猜测与失误参数都比模拟研究中给出的高不少，因此，模拟研究的结果可能不具备现实场景的可推广性。建议增加一个来自现实题库，或者根据现实题库模型生成的项目参数来进行模拟研究；

回应：非常感谢专家的提问。根据专家意见，我们已经修改了参数，选取了不同质量的题目进行模拟（见第 5 部分蓝色字体）。同时，我们还增加了实测数据，比较最简完备 Q 矩阵与可达阵的分类能力，研究结果发现，最简完备 Q 矩阵与可达阵的识别重复率达 90%以上，且最简完备Q矩阵识别的知识状态类型更多一些，说明题少的最简完备Q矩阵不逊于可达阵（见第 6 部分蓝色字体）。

意见 2：可以增加应用场景的说明，甚至根据应用场景来设计模拟研究。本文只是指出，新方法使用于多级计分，然后提出新方法可以提高测验效率，因此本人推断，这个方法会有助于“认知诊断题库建设以及短题长的自适应考试选题”这两个应用。但是这个只是本人的理论假设，需要模拟研究或者作者提供完整的应用场景。我想这样的工作会大大提升本文新方法的价值意义。

回应：非常感谢专家的建议。据此，在模拟研究中，我们增加了长测验与短测验情况，适用于终结性评估和过程评估。同时，我们增加了实测数据的研究，考察的是五年级数学中的行程问题，后续我们会进一步地考虑最简完备 Q 矩阵运用于自适应测验中（见第 5，6 部分蓝色字体）。

第二轮

审稿人 1 意见：

论文《基于可达阵的多级评分最简完备 Q 矩阵的设计》经过一轮修改后，作者较好地回答了审稿人的相关问题，论文质量得到较大提升。不过研究依旧存在以下问题：

回应：非常感谢专家对第一轮修改内容的肯定，对专家提出的意见和建议，我们将尽全力仔细修改文章，以期达到专家的要求。

意见 1：摘要当中的第 3 个结果“… 其判准率仍不低于于可达阵”中的‘于于’重复。

回应：非常感谢。我们已经删除重复的文字，并对全文进行了反复校对，尽量避免类似错误的发生。

意见 2：2.1 章节中，收敛结构的最简完备 Q 矩阵含有多个，如此，在测验设计时如何处理这多个最简完备 Q 矩阵？是任选一个最简完备 Q 矩阵？还是使用其它方法？

回应：感谢专家的提问。虽然收敛结构的最简完备 Q 矩阵有多个，但它们的功能是一样的。因为所选的列对应不同的分支且这些分支是并列关系，也即这几个最简完备 Q 矩阵的结构

相似，故这几种最简完备 Q 矩阵可以任选，或者如果测验设计含多个最简完备 Q 矩阵，为了避免曝光率，这些矩阵甚至可以混合使用（见 2.1 中蓝色字体）。

意见 3：作者在进行研究设计时，是如何操纵结构化最简完备 Q 矩阵、非结构化最简完备 Q 矩阵以及可达矩阵这三者的？在数据生成阶段操纵，还是在参数估计阶段？抑或是两个阶段都用到了？

回应：感谢专家的提问。根据实际需求，对于给定的属性及其层级关系，若测验必须考虑属性层级关系，则设计结构化最简完备 Q 矩阵或可达矩阵；若测验可以不考虑属性层级关系，则设计非结构化最简完备 Q 矩阵。所以，结构化最简完备 Q 矩阵、非结构化最简完备 Q 矩阵以及可达矩阵是在施测之前就设计好的。然后施测，根据所得 ORP，挑选合适的认知诊断模型进行诊断分类。

意见 4：作者所提出的最简完备 Q 矩阵确定过程涉及较多的步骤，为使后续的研究者可使用作者提出的最简完备 Q 矩阵方法，建议作者将相应的代码和程序给予公开。

回应：非常感谢专家的意见和建议。由于篇幅有限，代码不在文章列出，若拙文有幸能够发表，读者可向我们索取代码（见 2.2 蓝色字体）。

意见 5：模拟研究一中的研究设计和研究结果之间的对应关系不明确。在研究设计中作者操纵了长和短的测验长度，但在研究结果中，审稿人并不清楚哪些条件属于长测验条件，哪些又属于短测验条件。

回应：感谢专家的提问。文中 5.1.1 和 5.1.2 部分均提到长测验和短测验设计及其项目参数的设置，5.1.3 分别给出了长测验和短测验对应的模拟结果，长测验结果用表及文字进行说明，短测验结果用图和文字进行说明（蓝色字体为长测验内容和绿色字体为短测验内容）。

意见 6：模拟研究中的测验 Q 矩阵似乎没有呈现，如此审稿人难以了解测验 Q 矩阵的具体信息。

回应：感谢专家的提问。在第 5 部分的模拟研究中的第一段文字，其中蓝色字体“研究从属性层级结构（如图 1 和附录）……，考察（含）结构化/非结构化最简完备 Q 矩阵、可达阵等三种认知诊断测验的分类效果”提到了测验 Q 矩阵的具体信息。图 1 中的结构化最简完备 Q 矩阵在 2.1 中已经列出，其他属性的结构化最简完备 Q 矩阵见附录。由于非结构化最简完备 Q 矩阵较多，故在模拟中，随机挑选，取 PMR 和 MMR 的平均值作为最后结果（详见蓝色字体）。

意见 7：从 5.1.2 章节中的测验 Q 矩阵设计来看，测验 Q 矩阵似乎只包含一个可达矩阵，但很可能包含多个结构化/非结构化（最简）完备 Q 矩阵（后面简称‘完备 Q 矩阵’）；由于测验 Q 矩阵包含更多的完备 Q 矩阵，如此，完备 Q 矩阵下的 PMR 和 MMR 高于可达矩阵下的 PMR 和 MMR 似乎也在情理之中。因此，作者在得出完备 Q 矩阵的结果优于可达阵的结论时，应当谨慎。

回应：非常感谢专家的意见和建议，原先拙文的确没有清楚阐述这个问题。从 IRP 的角度，与可达阵一样，1 个最简完备 Q 矩阵能将 KS 完全区分，然而，在实际应用中，由于列数小于可达阵且存在测量误差，使得 1 个最简完备 Q 矩阵的判准率（PMR 或者 MMR）不一定高于（甚至抵不上）可达阵（见第 5 部分第一段开头蓝色字体）。通过模拟研究发现，在不同条件下，多个最简完备 Q 矩阵要达到一定的量，其判准率才会高于可达阵，即当多个最简完备 Q 矩阵组合使得其列数与可达阵相同或者略少时，最简完备 Q 矩阵判准率均高于可达

阵，其原因或许是重复测量使得误差缩小（见 5.1.3 、5.2.2 和 7.2 部分蓝色字体）。

意见 8：部分语句表述过于拗口，如 5.1.2 章节中的“采用最大后验估计方法（Maximum A Posteriori, MAP）估计被试，其中认知诊断模型为 RP-DINA 模型，评分方式为本文评分方式。”这三句话无法有效地组成一句表述清晰且连贯的句子。建议作者仔细通读全文，避免出现此类表述不完善的语句。
回应：非常感谢专家提出宝贵意见。我们对全文进行了仔细地修改（见 5.1.2 的（3）和（4）），希望本次修改能够达到您的要求。

意见 9：审稿人对于表 4-表 6 中的结果呈现有些看不懂。以表 4 中线形结构为例，为何‘结构化完备 Q 矩阵’在 1(1)、3(3)和 6(6)三种情况下均有 PMR 和 MMR，而另外二种矩阵则没有。另外，1(1)、3(3)和 6(6)的括号中数值究竟表示什么意思？1(1)中的(1)表示只有一个题目吗？如果是的话，为何只有一个题目？
回应：感谢专家的提问。这些 1(1)、3(3)和 6(6)括号外的数据描述有多少个最简完备 Q 矩阵，括号内的数据为这些最简完备 Q 矩阵的总列数。其中线型非结构化完备 Q 矩阵中没有提供 1(1)的 PMR 和 MMR 是因为线型结构的结构化最简完备 Q 矩阵只有 1 列（见 2.1 中线性结构部分），即（111111），而其对应单位阵中的列为（000001），根据非结构化最简完备 Q 矩阵来自于夹在两者之间的列，同时又必须满足每行至少有一个 1，这样的列是不存在的，故而没有提供相关数据。由于属性个数为 6，故可达阵的列数为 6 列，不存在 1(1)、3(3)的情况，也就不存在相关数据。

审稿人 2 意见：作者已经很好的回答了我的意见，推荐发表。
回应：非常感谢专家对论文肯定。

编委意见：这篇论文经过两位审稿人与作者的互动修改，达到了发表水平，建议发表。

主编意见：同意外审和编委意见，建议录用。