《心理学报》审稿意见与作者回应

题目：多级评分聚类诊断法的影响因素
作者：康春花，任平，曾平飞

第一轮

审稿人 1 意见：文章研究了各种因素对 GRCDM 模型的影响，有一定的研究意义。并且还有几个问题，请作者回答

意见 1：文中表 4 中，第 3 列是指改变的项目编号还是指改变的项目数量，如果是指编号，则是每种 Q 矩阵中只有一题存在属性向量误设的情况，请考察各种 Q 矩阵中存在不同个项目属性向量误设的情况，比如分别考察 1,2,3 个项目误设时的情况。

回应：谢谢专家的问题，专家的意见非常中肯。文中表4中，第3列是指改变的项目编号。在Q矩阵误设对判准率的影响方面，我们主要参考了De la Torre（2008）、Rupp & Templin（2008）和涂冬波，蔡艳，戴海琦（2012）等关于这方面的研究，见图1。在De la Torre（2008）的设计中，每种实验条件改变的项目数也是一个，所不同的是，他改变的类型中涉及到的属性稍多点。而在本研究中，由于属性数目只有4个，如果改变的题目数太多或属性数太多，必然导致对线型和收敛型不公正的局面，因其包含的考核模式本身就很少。

喻晓锋，罗照盛等人（2015）“基于作答数据的模型参数和Q矩阵联合估计”的研究中，在DINA模型的前提下，研究项目为20个，考察的属性个数分别是3、4和5个，初始Q矩阵中分别存在3、4 和5个属性界定错误的项目。结果表明：联合估计算法能在错误的初始Q 矩阵基础上以很高的概率得到正确的Q矩阵。在此研究中，由于用的是DINA模型，不考虑属性层级，可以设置较多的项目个数（20个），并且在属性数目相同的前提下同时考虑3、4、5个属性错误界定的情况，保证了研究前提的一致。而在我们的这个研究中，由于考虑了属性层级关系，Q矩阵是由属性层级导出的，如果要保持每个Q矩阵中的题目数量一致，必然线型要扩大2倍，则其包含的R矩阵又多了一个，而丁树良等人的研究表明，Q矩阵中包含的R矩阵个数会影响判准率，这又必然导致结果归因的困难与混淆。本研究中，尽管采用了传统的设置方法，但其研究结果还是发现了一些有意义的结论。

然而，专家的建议是非常有意义的，未来研究可以考虑在没有属性层级的前提下，采用喻晓锋，罗照盛等人（2015）等人关于Q矩阵误设和联合估计的方法，进一步考察GRCDM的稳定性与灵敏性。关于这一点，我们会在研究不足和展望中补充进去。以上是我们对此问题的看法，不知是否妥当，请专家审阅。

De La Torre, J. (2008). An Empirically Based Method of Q‐Matrix Validation for the DINA Model: Development and Applications. *Journal of Educational Measurement, 45*(4), 343-362.

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.

涂冬波, 蔡艳, & 戴海琦. (2012). 基于DINA模型的Q矩阵修正方法. *心理学报*(04), 558-568.

喻晓锋, 罗照盛, 秦春影, 高椿雷, & 李喻骏. (2015). 基于作答数据的模型参数和 Q 矩阵联合估计. *心理学报,47*（ *2*），273-282.

实验条件	改变的项目	改变类型
1	1	$\alpha_1 \to 0, \alpha_2 \to 1$
2	1	$\alpha_2 \to 1$
3	11	$\alpha_1 \to 0, \alpha_3 \to 1$
4	11	$\alpha_1 \to 0$
5	11	$\alpha_3 \to 1$
6	21	$\alpha_1 \to 0$
7	21	$\alpha_1 \to 0, \alpha_2 \to 0$
8	21	$\alpha_1 \to 0, \alpha_4 \to 1$
9	21	$\alpha_1 \to 0, \alpha_2 \to 0, \alpha_4 \to 1$
10	21	$\alpha_1 \to 0, \alpha_2 \to 0, \alpha_4 \to 1, \alpha_5 \to 1$
11	1 11 21	$\alpha_2 \to 1$ $\alpha_2 \to 0, \alpha_3 \to 0$ $\alpha_1 \to 0, \alpha_4 \to 1$

资料来源：de la Torre (2008)

图 1 de la Torre 关于 Q 矩阵错误的设定

意见 2：本文中不同属性层级关系下的 Q 矩阵项目数不同，但是放在一起进行比较，因为项目数量对于认知诊断测验非常重要，这一点差别无法忽视。

回应：感谢专家的提问。关于属性层级关系对判准率的影响方面，已有研究(田伟, & 辛涛，2012；祝玉芳, & 丁树良，2009；涂冬波，蔡艳，戴海琦, & 丁树良，2010；罗欢，丁树良，汪文义等，2010，以及其它相关研究)都是采用类似范式，即由属性层次结构得到 R 矩阵，再由 R 矩阵直接导出简化 Q 阵，然后基于简化 Q 阵进行后续研究，并未考虑到尽管属性个数相同，但由于层级关系不同，导致各个简化 Q 阵题目数不同对结果的影响，而是在此基础上直接下结论，说判准率受属性层级紧密度的影响，关系越紧密，判准率越高。本文遵照前人研究范式，从层级关系得到 R，由此导出简化 Q，在此基础上进行后续研究，并且还有一个改进，因无结构型题目太多，有 64 题，本文在保证 Q 包含 R 的基础上，参照罗欢等人(2010)的研究，减缩为 22 题（前人很多研究并未减缩），这样不至于无结构型的题目数与其它相差太大。由此，本研究的设计范式与前人相当。此外，本文的 GRCDM 是一种非参数的方法，该方法采用能力向量作为指标，实际上能力向量是对属性合分向量的标准化，即是为消除属性被测次数不同而导致的额外影响，因此，本研究所下结论应该是较为妥当的。

要排除题目个数的影响，在不同层级结构下很难做到。因为层级紧密度不一样，必然导致 R 不一样，由此导出的简化 Q 不一样。要使得题目数平衡，有两种做法，一是把无结构型和发散型的题目数大大减缩，得到其题目数与线型和收敛型相当，但这样做会导致减少后的无结构型和发散型的 Q 阵不一定包含了 R 阵，从而导致理想反应模式与知识状态不能一一对应，由此导致乱判。第二种做法是把线型和收敛型的题目数增加，比如在原有基础上增加 3 倍，使得其题目数与无结构型和发散型相当，但这样做也会有问题。因为在原有基础上增加 3 倍，相当于 Q 阵包含了 3 个 R 阵，研究表明 Q 阵中包含的 R 阵越多，判准率越高(丁树良，毛萌萌，汪文义等，2012；丁树良，汪文义，杨淑群，2011)。此时用只包含 1 个 R 的无结构型和发散型与包含了 3 个 R 的线型与收敛型比较，明显是不对等的，我们很难判断判准率的提高是由于题目数的影响还是多个 R 的影响。当然，还有另一条思路，就是在固定某一层次结构的基础上，从同一简化 Q 阵中抽取不同题目数的组合模式，来考察题目数对判准

率的影响，但此时就很难对不同的层级结构进行比较了。因此，关于题目数对判准率的影响，这是一个非常有趣的也是值得继续深入研究的方向，后续研究中我们会继续进一步探讨此问题。感谢专家的宝贵意见。

丁树良, 毛萌萌, 汪文义, 罗芬, & Ying, C. (2012). 教育认知诊断测验与认知模型一致性的评估. *心理学报, 44*(011), 1535-1546.

丁树良, 汪文义, & 杨淑群. (2011). 认知诊断测验蓝图的设计. 心理科学(02), 258-265.

罗欢, 丁树良, 汪文义, 喻晓锋, & 曹慧媛. (2010). 属性不等权重的多级评分属性层级方法. *心理学报 42*(04), 528-538.

田伟, & 辛涛. (2012). 基于等级反应模型的规则空间方法. *心理学报, 44*(1), 249-262.

涂冬波, 蔡艳, 戴海琦, & 丁树良. (2010). 一种多级评分的认知诊断模型: P-DINA 模型的开发. *心理学报 42*(10), 1011-1020.

祝玉芳, & 丁树良. (2009). 基于等级反应模型的属性层级方法. *心理学报 41*(03), 267-275.

意见 3：缺乏比较研究，建议增加本文模型的表现与其它已发表的模型（APM，心理学报都有发表相关的模型）的比较。

回应：感谢专家非常中肯的意见。目前，多级评分的认知诊断模型已经不少，如 Bolt 和 Fu（2004）的多级 Fusion 模型、祝玉芳等人（2009）和罗欢等人（2010）的 GRM-AHM、涂冬波等人（2010）的 P-DINA 模型、田伟和辛涛（2012）的多级规则空间方法、李娟，丁树良和罗欢（2012）的基于等级反应模型的广义距离判别法（GRM-GDD）、张淑梅，包钰和郭文海（2013）多级评分的广义认知诊断模型（GP-DINA）、Sun, Xin, Zhang, &de la Torre(2013) 提出的多级计分的广义距离判别方法（GDD-P）等。关于 GRCDM 与其它方法的比较，我们会结合第二位专家的意见一起回复，因为两位专家的意见基本一致但稍有不同，我们在下面对第二位专家的第 4 个问题的回复中一并阐述，这里就不赘述。

B o ltD, Fu J. B. (2004) . A polytomous extension of the fusion model and its Bayesian parameter estimation. Paper presented at NCM E, San Diego, USA.

李娟, 丁树良, & 罗芬. (2012). 基于等级反应模型的广义距离判别法. *江西师范大学学报(自然科学版)*(06), 636-639.

罗欢, 丁树良, 汪文义, 喻晓锋, & 曹慧媛. (2010). 属性不等权重的多级评分属性层级方法. *心理学报*(04), 528-538.

Sun, J., Xin, T., Zhang, S., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement*,37(7)，503-521.

涂冬波, 蔡艳, 戴海琦, & 丁树良. (2010). 一种多级评分的认知诊断模型:P-DINA模型的开发. *心理学报* (10), 1011-1020.

田伟, & 辛涛. (2012). 基于等级反应模型的规则空间方法. *心理学报*(02), 249-262.

张淑梅, 包钰, & 郭文海. (2013). 一种多级评分的广义认知诊断模型. *心理学探新*(05), 444-450.

祝玉芳, & 丁树良. (2009). 基于等级反应模型的属性层级方法. *心理学报*(03), 267-275.

审稿人 2 意见：

意见 1：本文采用的研究方法和模型与前人研究（康春花等，2015）完全相同，只是增加了

几个变量和水平。文章的创新性有所不足。

回应：感谢专家的提问。通过查阅资料，康春花等人在 2015 年有两篇研究，一是康春花，任平（2015）关于“聚类诊断分析法诊断正确率的影响因素”；二是康春花，任平，曾平飞（2015）关于“非参数认知诊断方法：多级评分的聚类分析”。前者是关于 0-1 评分聚类诊断法的影响因素研究，是一个较基础的小模拟研究，探讨了属性个数、样本容量和层级关系对 0-1 聚类诊断方法的影响，结果表明：0-1 聚类诊断方法对样本容量无依赖，属性个数与层级关系存在交互作用，当属性个数较多时，0-1 聚类法较适合分散型结构。后者的侧重点是把 0-1 评分的聚类诊断法拓展到多级评分，并通过模拟（样本容量、属性层级和失误率）和实证研究探讨多级评分聚类诊断法（GRCDM）的性能，研究表明：GRCDM 的性能与 0-1 评分方法性能类似，即对样本容量无依赖，较适合松散型结构，随失误率增大判准率会有所下降（这与参数方法的结果也一致）。可见，前者只是对 0-1 评分方法的探讨，后者是把 0-1 方法拓展到了多级评分。然而，两者在影响因素上的探讨都只是设置了一些基本变量，并没有从被试、测验等角度全面的考察其性能。因此，为进一步全面的考察 GRCDM 的性能，才有了本研究，所以本研究是与以往研究不同的，是以往研究的深入和继续。

本研究采用的模型是与康春花，任平，曾平飞（2015）的相同，即是对 GRCDM 性能的深入探讨，但研究目的、研究内容、侧重点和研究发现都是不一样的。本研究试图在康春花，任平，曾平飞（2015）某些研究结果的基础上，参照参数诊断模型一些有意义的做法，对 GRCDM 这种非参数诊断方法的稳定性和灵敏性进一步进行评估。因为目前，关于非参数诊断方法的深入研究还是比较少的，所以本研究的目的比较简单，就是想深入了解 GRCDM 这种非参数诊断方法的性能，抛砖引玉，供专家审阅和读者讨论，以丰富认知诊断方法的理论和实践探讨。所以，诚如专家所言，在创新性方面有所不足，但确实是以往研究的继续和深入。

康春花等（2015）的主要目的是将 0-1 评分的聚类分析法拓展到多级评分(Grade Response Cluster Diagnostic Method , GRCDM)，因此如何拓展是其重点之一。此外，初步探讨了一些基础变量，如样本容量、失误率及属性层级对其判准率的影响，并进一步考察了其在实践中的表现。失误率和属性层级是惯常的考察因素，而加入样本容量是为了说明非参数方法对人数多少无依赖的特点。康春花等人（2015）的研究结果表明：GRCDM 在模拟和实践情境中均有很高的判准率，且对样本容量及属性层级紧密度依赖较小，可适用于小型测评等特征，这在一定程度上体现出非参数方法的优势。康春花等人（2015）得到了关于 GRCDM 的一些初步结论，然而，在其它参数诊断模型已关注的方面，如属性数目、被试分布、属性层级误设、Q 矩阵误设等，GRCDM 并未涉及。

针对目前关于非参数方法的研究还尚粗浅的现状，能否借助参数诊断模型的已有成果，探索 GRCDM 的影响因素，深入考察 GRCDM 的优势和性能，丰富非参数方法研究，是本研究的初衷。为此，我们在梳理参数诊断模型相关研究的基础上，将影响模型判准率的因素概括为三个方面：一是与诊断测验相关的因素，如属性层级关系、Q 矩阵、属性个数、题目数量（测验长度）等；二是与被试相关的因素，如被试能力分布、样本容量、失误率等；三是模型的选择，如模型与数据是否拟合，或模型与题目特征是否吻合（问题解决时属性之间的补偿性）。康春花等（2015）的研究主要考虑了属性层级、样本容量和失误率三个因素。在失

误率方面得到了与参数诊断模型一致的结论，即失误率越高判准率越低；在属性层级方面，得到了与参数诊断模型不同的结论；而在样本容量方面，GRCDM 具有无依赖的特性。因此，为较全面的考察 GRCDM 的性能，结合康春花等（2015）已有结论，在本研究中，我们将失误率和样本容量固定，而留给更多的空间，来探讨其它影响因素，及其与层级关系的交互作用，以期更加丰富和深入地扩展 GRCDM 的研究。

为此，本研究从测验和被试两个层面，通过精心设计的三个模拟研究，探讨了属性数目、被试分布、属性层级关系、属性层级误设和 Q 矩阵误设等 5 个因素对 GRCDM 的影响，试图尽量全面的考察 GRCDM 的性能，推动非参数诊断方法的研究。本研究是以往研究的深入和扩展，通过三个模拟研究，得到一些有意义的结论，这些结论有些是非参数方法所特有的，有些是与参数方法共有的。如，所得结果表明：(1)GRCDM 对属性数目无依赖，随属性数目的增多判准率反而增高。这与参数研究的结果不同，参数方法一般对属性数目有要求，一般在 5,6 个左右，当多于 7 个时，判准率会急剧下降；(2)被试能力分布对 GRCDM 判准率高低无影响，这也是与参数方法不同之处，更多的体现了非参数方法对被试能力分布无要求的优势；(3) GRCDM 对属性层级误设的灵敏性（稳定性）因层级关系的不同而不同，无结构型和发散型时，“属性层级关系错乱”的判准率降幅最大。(4)Q 矩阵误设对 GRCDM 的影响因层级关系而异，收敛型和发散型受影响较小，无结构型和线型的判准率在属性既冗余又缺失时降幅最大。这一点与参数方法所得结论有相同也有不同之处，具体内容见文中阐述。总之，通过此研究，我们更多地了解了 GRCDM 的特性，了解了非参数方法的优势及其与参数方法的区别，为认知诊断评估的理论研究和实践中的模型和方法选用提供了有用信息。

综上，我们认为本研究与康春花等（2015）无论是从研究目的、内容结构、还是研究发现等方面都存在较大的差异，本研究是前者的继续和深入，所得结论对于研究非参数诊断方法具有一定的参考意义。以上是我们对此问题的看法，不知妥否，请专家审阅。

康春花, 任平, & 曾平飞. (2015). 非参数认知诊断方法: 多级评分的聚类分析. *心理学报, 47*(8), 1077-1088.
康春花, & 任平. (2015). 聚类诊断分析法诊断正确率的影响因素. *中国考试*(2), 25-32.

意见 2：本文中设定的 10%的失误率是否过低？作者所得到的较高的判准率是否因为失误率较低导致？

回应：感谢专家的提问。专家的意见很中肯。在已有关于失误率对判准率的影响研究中，都得到一致性的结论，即失误率越高则判准率越低，这在康春花等人(2015)的研究中也一样得到了证实。然而，本研究所关注的重点并不在失误率与判准率之间的关系（这个在康春花等人（2015）中已经研究过，结果表明即使在失误率为 20%时，其判准率也保持了较高的比率），而在于探讨其它因素对 GRCDM 的影响，因此固定了判准率。

关于失误率的设定，我们查阅了一些相关文献，发现 10%的失误率算是个中等水平，见表 1。表 1 表明，除了在涂冬波，蔡艳，戴海琦（2012）和 Sun, J., Xin, T., Zhang, S., & de la Torre, J. (2013)的研究中（表中标黄部分），10%的失误率至少是个中等水平。在 Sun 等人（2013）的研究中，他们所提出的 GDD-P 在 30%失误率时，其模式判准率至少都在 0.8222 以上，只

有到 40%时，其判准率才突降至 0.6083。GDD-P 总体来说具有很高的判准率，然而其判准率也随失误率的增大而下降，这与以往研究的结论也是一致的。

在本研究中，其目的不是探讨失误率与判准率之间的关系，而是讨论其它因素对 GRCDM 的影响，因而把失误率固定在了某一个水平。当然，在失误率的不同水平，其它因素对 GRCDM 的影响是始终保持一致呢？还是会存在交互作用，这个在本研究中并未涉及。我们的猜测是应该保持一致的，是否如此，值得进一步加入失误率这个变量进行探讨，后续研究会继续关注。请专家审阅。

表 1 关于失误率的设定

文献	失误率设定
祝玉芳，丁树良（2009）	2%、5%、10%、15%
罗欢，丁树良，汪文义等（2010）	5%、10%、15%、20%
田伟，辛涛（2012）	2%、5%、10%、15%
涂冬波，蔡艳，戴海琦，丁树良（2012）	2%、5%、10%
涂冬波，蔡艳，戴海琦（2012）	5%、10%、15%、20%、25%
涂冬波，蔡艳，戴海琦（2013）	2%、5%、10%
李娟，丁树良，罗欢（2012）	2%、5%、10%、15%
Sun, J., Xin, T., Zhang, S., & de la Torre, J. (2013)	10%、20%、30%、40%

李娟, 丁树良, & 罗芬. (2012). 基于等级反应模型的广义距离判别法. *江西师范大学学报(自然科学版)*(06), 636-639.

罗欢, 丁树良, 汪文义, 喻晓锋, & 曹慧媛. (2010). 属性不等权重的多级评分属性层级方法. *心理学报*(04), 528-538.

田伟, & 辛涛. (2012). 基于等级反应模型的规则空间方法. *心理学报, 44*(1), 249-262.

涂冬波, 蔡艳, & 戴海琦. (2012). 基于DINA模型的Q矩阵修正方法. *心理学报*(04), 558-568.

涂冬波, 蔡艳, & 戴海琦. (2013). 几种常用非补偿型认知诊断模型的比较与选用:基于属性层级关系的考量. 心理学报, 45(02), 243-252.

涂冬波, 蔡艳, 戴海琦, & 丁树良. (2012). 一种多策略认知诊断方法:MSCD方法的开发. *心理学报*(11), 1547-1553.

祝玉芳, & 丁树良. (2009). 基于等级反应模型的属性层级方法. *心理学报 41*(03), 267-275.

Sun, J., Xin, T., Zhang, S., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement*,37(7)，503-521.

意见 3：作者设定的失误种类单一，只有向临近得分滑动的失误。基于此得到的结果是不是局限太大？已有研究中采用了更加多元的多级计分失误方法。比如本来应该得 3 分的，按照作者的失误生成方法，只可能得 2 分或 4 分，但现实情境中会不会有 1 分或 5 分的可能？

回应：谢谢专家的意见。专家的意见非常中肯。本研究关于失误种类的设计，是参照以往研究（田伟，辛涛，2012；祝玉芳，丁树良，2009；罗欢，丁树良，汪文义等，2010；涂冬波，蔡艳，戴海琦，丁树良，2010；李娟，丁树良，罗欢，2012）进行的。在这些关于多级计分认知诊断模型的研究中，其分数失误的方法都是采用向临近得分滑动的结果，所以本研究也及采用了以往的研究范式，便于做相关结果的比较。

然而，正如专家所言，这种加 1 分减 1 分的范式，现在看来是太单一了，应该采用与现实情境较吻合的多元失误方式。比如，对于满分本身较低的项目，可以采用向临近得分滑动的方式，而对于满分较高的项目，则可以采用左右滑动 1 分、2 分甚至 3 分的范式。在张淑梅，包钰和郭文海（2013）的研究中，他们就是采用多元滑动的方式设计失误种类的。他们提出了一种有多个潜变量多个滑动参数的多级评分认知诊断模型——GP-DINA，采用项目滑动矩阵的概念，并利用 EM 算法估计滑动矩阵，从而得到模型的参数估计值及被试的属性掌握模式，研究表明线型、收敛型、发散型、无结构型和独立型均有较高的判准率。受专家意见和张淑梅等人（2013）的启发，尽管关于 GP-DINA 中用 EM 算法估计滑动矩阵的技术，我们现在还没有掌握，但其关于项目滑动矩阵这种多元化失误分数的设定范式，是值得我们借鉴的。因此，我们后续的其它研究采用了这个思想。而关于此研究，我们会参照这篇研究，在研究不足与展望中进行阐述。请专家审阅。

李娟, 丁树良, & 罗芬. (2012). 基于等级反应模型的广义距离判别法. *江西师范大学学报(自然科学版)*(06), 636-639.

罗欢, 丁树良, 汪文义, 喻晓锋, & 曹慧媛. (2010). 属性不等权重的多级评分属性层级方法. *心理学报*(04), 528-538.

涂冬波, 蔡艳, 戴海琦, & 丁树良. (2010). 一种多级评分的认知诊断模型:P-DINA模型的开发. *心理学报* (10), 1011-1020.

田伟, & 辛涛. (2012). 基于等级反应模型的规则空间方法. *心理学报*(02), 249-262.

张淑梅, 包钰, & 郭文海. (2013). 一种多级评分的广义认知诊断模型. *心理学探新*(05), 444-450.

祝玉芳, & 丁树良. (2009). 基于等级反应模型的属性层级方法. *心理学报*(03), 267-275.

意见 4：为了证明 DRCDM 是个优秀的认知诊断方法，是否需要将其表现与其他模型或方法进行比较？如果列出相同条件下，DINA 模型，RSM,AHM 等认知诊断方法的结果，加以比较，相信会更有说服力。特别是在属性层级设定错误的情况下，读者对于其他方法在此情况下的表现并不一定有一个清晰的概念，增加对比有助于增加结果的可信度。

回应：感谢专家的指点，专家的意见非常中肯。目前，多级评分的认知诊断模型已经不少，如 Bolt 和 Fu（2004）的多级 Fusion 模型、祝玉芳等人（2009）和罗欢等人（2010）的 GRM-AHM、涂冬波等人（2010）的 P-DINA 模型、田伟和辛涛（2012）的多级规则空间方法(GRM-RSM)、李娟，丁树良和罗欢（2012）的基于等级反应模型的广义距离判别法（GRM-GDD）、张淑梅，包钰和郭文海（2013）多级评分的广义认知诊断模型(GP-DINA)、Sun, Xin, Zhang, &de

la Torre(2013)提出的多级计分的广义距离判别方法（GDD-P）等。关于 GRCDM 与其它多级评分模型的比较，在康春花，任平，曾平飞（2015）已经有部分阐述，如表 2 所示。表 2 表明 GRCDM 无论在何种条件下，其模式判准率和属性判准率都要优于其它方法，由于康春花等（2015）已作阐述，本文就不再赘述。

表 2 GRCDM 与其它多级评分诊断模型判准率的比较

方法	分析水平	2%	5%	10%	15%
GRCDM	模式判准率	.997	.992	.978	.958
	属性判准率	1.000	.999	.997	.994
GRM-AHM-A	模式判准率	.953	.914	.897	.836
	属性判准率	.993	.987	.984	.974
GRM-AHM-B	模式判准率	.904	.777	.600	.445
	属性判准率	.955	.896	.813	.733
GRM-AHM-LL	模式判准率	.978	.942	.898	.850
	属性判准率	.994	.985	.974	.959
GRM-RSM	模式判准率	.957	.882	.789	.644
	属性判准率	.990	.974	.953	.922

资料来源：康春花，任平，曾平飞（2015）

按照本研究的研究设计及专家的意见，我们在讨论部分增加了与本研究相似条件下的结果比较，主要包括 3 部分：（1）是属性个数变化时，GRCDM 与其它模型的判准率比较；（2）是被试知识分布状态不同是，GRCDM 结果与其它模型的比较；（3）是属性层级误设时 GRCDM 与其它模型结果的比较。

首先，在不同属性个数条件下，相关多级评分认知诊断方法的结果比较见表 3。由表 3 可以看出，与参数方法不同的是，GRCDM 判准率不仅不随属性数目的增多而减低，反而呈增高的趋势，并且，在属性个数相当甚至较多的情况下，其判准率要高于 P-DINA、GRM-GDD、GRM-AHM-A、GRM-AHM-B、GRM-RSM 和多级 Fusion 等方法，但略微低于 GDD-P 和 GP-DINA 两种方法。可见，在属性数目较多样本容量又较少的情况下，较适于选用 GRCDM 作为分类方法，但如果样本容量较大，则 GDD-P 和 GP-DINA 也是不错的选择。

表 3 GRCDM 与其它多级评分诊断模型在不同属性个数的判准率(%)

判准率	GRCDM			P-DINA					GRM-GDD	GRM-AHM-A
	4	7	9	4	5	6	7	8	7	7
MMR	97.85	99.56	99.85	98.9	97.4	96.3	93.3	90.5	----	----
PMR	93.25	97.51	98.78	95.7	88.4	80.7	66.0	52.9	96.0	88.0

续表 3

判准率	GRM-AHM-B	GRM-RSM	GDD-P	GP-DINA	多级 Fusion
	7	7	7	6	4
MMR	----	83.1	99.9	99.9	----
PMR	33.4	48.3	99.6	99.9	94.9

其次，在被试知识状态不同条件下的比较，见我们对第 7 个问题的回复，不再赘述。

最后，属性层级误设时，GRCDM 与其它方法的比较。在这个比较方面，由于表 3 的多级评分诊断模型，都没有考虑属性层级误设及 Q 矩阵误设方面的研究，所以本研究考虑与涂冬波，蔡艳和戴海琦（2013）关于 0-1 计分模型的比较，从而可以大致看出 GRCDM 在属性层级误设时的稳定性，结果见图 2（图中纵坐标单位为%）。图 2 列出了 GRCDM 与其它方法在属性层级误设时的降幅，从图中可以看出，GRCDM 无论在各种属性层级误时的降幅还是总体平均降幅都比 RSM、AHM-A、GDD 要小很多，但比 DINA-HC 稍高。由此，我们可以认为 GRCDM 在层级误设时的判准率还是比较稳定的，之所以比 DINA-HC 模型要稍高点，是因为 DINA 模型族本身就是不考虑层级关系的模型。

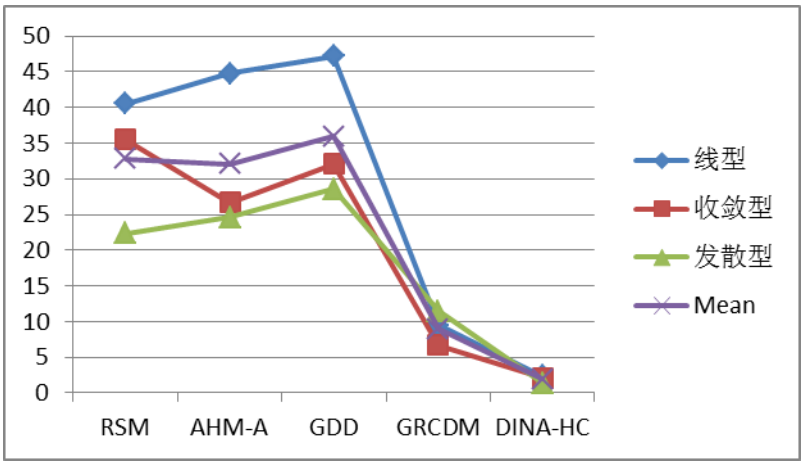

图2 属性层级误设时GRCDM与其它模型的判准率降幅比较

B o ltD, Fu J. B. (2004) . A polytomous extension of the fusion model and its Bayesian parameter estimation. Paper presented at NCM E, San Diego, USA.

李娟, 丁树良, & 罗芬. (2012). 基于等级反应模型的广义距离判别法. *江西师范大学学报(自然科学版)*(06), 636-639.

罗欢, 丁树良, 汪文义, 喻晓锋, & 曹慧媛. (2010). 属性不等权重的多级评分属性层级方法. *心理学报*(04), 528-538.

Sun, J., Xin, T., Zhang, S., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement*,37(7)，503-521.

涂冬波, 蔡艳, 戴海琦, & 丁树良. (2010). 一种多级评分的认知诊断模型:P-DINA模型的开发. *心理学报* (10), 1011-1020.

蔡艳, 涂冬波, & 丁树良. (2013). 五大认知诊断模型的诊断正确率比较及其影响因素: 基于分布形态, 属性数及样本容量的比较. 心理学报, 45(011), 1295-1304.

田伟, & 辛涛. (2012). 基于等级反应模型的规则空间方法. *心理学报*(02), 249-262.

张淑梅, 包钰, & 郭文海. (2013). 一种多级评分的广义认知诊断模型. *心理学探新*(05), 444-450.

祝玉芳, & 丁树良. (2009). 基于等级反应模型的属性层级方法. *心理学报*(03), 267-275.

意见 5：表 3 中为何要提到实验次数？表 3 是不是表 2 的总结？

回应：感谢专家的问题。表 2 是各层级误设类型下，其具体错误层级的 MMR 及其降幅均值，是描述统计量值。为了看出层级误设类型对 MMR 降幅影响的差异，我们进一步对层级误设类型进行了单因素方差分析，结果发现层级误设类型对 MMR 降幅的影响达到了显著水平，表 3 则是多重比较的结果。表 3 中的试验次数是每种错误类型下的重复试验次数，按照 SPSS 的结果，在 Scheffe 方法中会自动统计并显示出来，可以表明每种条件下的均值是基于多少次试验条件下的结果，所以就列在表中没有删除。

表 3 试验次数对应的 sig 是 1，如何解释？

回应：感谢专家指出该问题。非常不好意思，这是我们排版的错误，sig 对于的 1 应该是在 subset 1 下面的，我们已经在文中进行了修订。表中：不同的 subset 下的值，存在显著差异，而在同一个 subset 下的值表明并不存在显著差异。所以表 3 结果表明：层级关系错误的降幅均值 0.245 显著高于层级关系颠倒（0.088）和无层级变为有层级（0.104），层级关系颠倒和无层级变为有层级的降幅均值又显著高于有层级变为无层级（0.006），但层级关系颠倒和无层级变为有层级两者的 MMR 降幅均值并无显著差异。

意见 6：Q 矩阵误设和属性层级关系误设是什么关系？为何需要考虑这两个条件？

回应：感谢专家的提问，专家的问题非常中肯。关于这个问题，我们是这样看的：已有研究表明属性层级误设和 Q 矩阵误设都会影响诊断方法或模型的判准率（De la Torre，2008；Rupp & Templin，2008；涂冬波，蔡艳，戴海琦，2012；Im & Corter，2011；喻晓锋, 罗照盛，秦春影等，2015)。

属性层级误设与 Q 矩阵误设存在于认知诊断过程的不同阶段。属性层级误设出现在属性之间逻辑关系的界定过程中，属性层级关系的正误直接导致的是认知诊断测验编制的质量。比如，专家界定的层级关系与学生的认知过程本身存在不一致（喻晓锋, 罗照盛，秦春影等，2015），此时的层级关系可能不合理，据此编制的测验可能就存在一定的问题。探讨属性层级关系误设对判准率的影响，说到底是为了考察据此编制的测验对学生属性掌握模式判准率的影响，关系到的是认知诊断方法或模型在测验的效度或认知模型的效度验证上的问题。因此，含层级关系的认知方法或模型一般都会考察属性层级误设对其诊断结果的影响。

Q矩阵误设一般考察的是各个题目所涉及的属性是缺失的还是冗余的。因此，Q矩阵误设存在于题目编写或题目所涉及属性的认定中，但此时属性层级关系可以是正确的，即项目的考核模式是正确的，只是在题目编写或属性认定时有可能产生错误。比如，在自上而下的测验编制中，根据正确的属性层级可以推导出正确的Q矩阵，但是在编写题目时，这个题目的属性认定是不是正确的，如果认定的属性不明确，则有可能出现题目属性缺失或冗余。此外，在后补性的测验中，由于测验本身不是因认知诊断评估自上而上设计的，则根据题目来界定Q矩阵时，也有可能导致对题目考察属性的判断错误，此时出现题目属性的冗余或缺失。因此，Q矩阵误设考察的是题目编写的效果或称项目效度。

因此，我们认为属性层级误设与Q矩阵误设发生阶段和对结果的影响机制是不一样的，属性层级误设会影响测验编制的质量，从而影响诊断效果，并且其影响相对少数几个项目中的属

性冗余或缺失来说，可能是更严重的，因为属性层级误设必然导致不止少数项目的属性冗余或缺失。而Q矩阵误设只是可能发生在局部题目或属性上。基于此，本研究设计了这样两个子研究，试图从整体和微观的角度同时考察GRCDM的性能。请专家审阅。

De La Torre, J. (2008). An Empirically Based Method of Q‐Matrix Validation for the DINA Model: Development and Applications. *Journal of Educational Measurement, 45*(4), 343-362.

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. Educational and Psychological Measurement, 71(4), 712-731.

齐冰. (2008). HCI 对认知属性层次结构构建失误的侦查研究 (Master's thesis, 江西师范大学).

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.

涂冬波, 蔡艳, & 戴海琦. (2012). 基于DINA模型的Q矩阵修正方法. *心理学报*(04), 558-568.

涂冬波, 蔡艳, & 戴海琦. (2013). 几种常用非补偿型认知诊断模型的比较与选用基于属性层级关系的考量. 心理学报, 45(2), 243-252.

喻晓锋, 罗照盛, 秦春影, 高椿雷, & 李喻骏. (2015). 基于作答数据的模型参数和 Q 矩阵联合估计. *心理学报,47*（*2*），273-282.

意见 7：被试知识状态分布对于认知诊断结果有何影响？是不是因为有的知识状态估计的准确，有的知识状态估计的不准确，改变被试知识状态的分布犹如改变加权平均数的权重一样，可以影响最终判准率的结果？与其讨论被试知识状态分布，是否不如讨论对于不同知识状态被试的判准率？

回应：感谢专家的问题。在本研究中，之所以会考虑被试知识状态分布这个因素，主要是受蔡艳，涂冬波，丁树良（2013）“五大认知诊断模型的诊断正确率比较及其影响因素：基于分布形态、属性数及样本容量的比较”这篇研究的启发。在这篇研究中，作者研究了知识状态分布（正态分布、正偏态分布、负偏态分布和均匀分布）对5大认知诊断模型（RSM、AHM、DINA、GDD、DINA-HC）判准率的影响，所得结果表明：

（1）5个模型诊断正确率会受到知识状态分布形态的影响；

（2）对于RSM而言，不同知识状态分布下整体诊断正确率由高到低的分别是：负偏态、正偏态、正态、均匀分布；

（3）对于AHM和DINA模型，整体诊断正确率由高到低的分别是：负偏态、正态、正偏态、均匀分布；

（4）对于GDD和DINA_HC模型，整体诊断正确率由高到低的分别是：负偏态、均匀、正态、正偏态分布；

（5）对于不同模型，知识状态的分布特征对不同模型诊断正确率的影响各不相同，但对5种模型而言，当知识状态为负偏态时，5 个模型的诊断正确率均达最高，这可能与负偏态时绝大部分被试为高能力(即掌握更多的属性数)有关；

（6）不论在何种知识状态分布下，诊断正确率最高的模型是DINA_HC和DINA模型(PMR达90%以上)，其次为GDD模型(PMR达85%以上), AHM (PMR达65%以上)诊断正确率一般，RSM的诊断正确率最低，且其PMR基本不足50%。

在这篇研究中，作者比较的模型都是参数或半参数模型，研究结果表明被试知识分布状态会影响模型的判准率，且影响方式随模型不同而异。于是，我们就想，为什么被试知识状态会影响判准率呢？这可能与作者使用的都是参数模型有关，因为参数方法本身除了对样本容量

有依赖，还对总体分布有要求。

然而，我们的这个研究，用的是非参数认知诊断方法，而非参数方法的优势是前提假设比较弱，即对样本容量无依赖（康春花，任平，曾平飞（2051）已经验证）、对总体分布无要求等。正是基于这种考虑，本研究才想要考察一下GRCDM这种非参数方法在被试群体不同知识分布状态下，其判准率会不会受到影响。我们的研究结果表明：被试知识状态的分布及其与其它因素的交互对GRCDM判准率的影响甚微，其效果量太小，所以几乎可以忽略，因为，在均匀和正态分布下，各属性层级的PMR分别为：99.10、98.53、99.14、99.40；99.31、98.67、99.14、99.40）。这个结果说明GRCDM这种非参数方法对被试知识状态分布无依赖，这是与涂冬波等（2013）参数方法得到的结果不同的。为什么会这样？专家的意见可能会是一个比较好的解释：对于非参数方法而言，因其对每种得分的人数多少并无要求（但参数方法有），所以改变被试知识状态的分布就犹如改变加权平均数的权重一样，对非参数方法的结果并无影响，这正是非参数方法的优势所在。所以，谢谢专家的意见和提醒，按照专家的意见，我们会在讨论中加入我们的阐述。不知我们的理解是否妥当，请专家指正！

康春花, 任平, & 曾平飞. (2015). 非参数认知诊断方法: 多级评分的聚类分析. *心理学报, 47*(8), 1077-1088.
康春花, & 任平. (2015). 聚类诊断分析法诊断正确率的影响因素. *中国考试*(2), 25-32.
蔡艳, 涂冬波, & 丁树良. (2013). 五大认知诊断模型的诊断正确率比较及其影响因素: 基于分布形态, 属性数及样本容量的比较. 心理学报, 45(011), 1295-1304.

以上是我们对两位专家所提问题的回应，再次感谢专家的意见和建议！

第二轮

审稿人 1 意见：

意见 1：我有一个问题想问作者的是：正文中表 6 的结果，作者对于相关的条件和背景并没有交待，是直接引用的数据还是自己模拟实验的结果，这个需要交待清楚。

回应：感谢专家的问题。正文表 6 的结果是直接引用已有多级评分模型相似条件下的研究结果。感谢专家指出该问题。我们已经在文中进行了说明，如下：为比较不同模型在属性数目变化时判准率的变化趋势，搜索已有研究相似条件下的 9 种多级评分模型的模拟结果进行描述（见表 6）。请专家审阅!

审稿人 2 意见：

意见 1：感谢作者细致的回答了审稿人的问题，文章的可读性有所提升，同意发表。

回应：感谢专家对我们工作的肯定。

以上是我们对两位专家意见的回复，请编委复审，也为您接下来的辛苦工作致谢！