《心理学报》审稿意见与作者回应

题目：非参数认知诊断方法：多级评分的聚类分析
作者：康春花　任　平　曾平飞

第一轮

感谢专家给出的宝贵意见，我们尽量遵照专家的意见进行修改。在意见回应部分，对专家的意见我们用红色标注，我们的回应是黑色字体；在文中，结合专家意见，我们增加、重写的地方，也用红色标注，以便专家审查。以下是对专家意见的回应。

第一位审稿专家意见以及相关的回应

本研究使用非参数的方法进行认知诊断，这是一个很好的尝试，采用的多级评分聚类方法也确实具有一定的优越性。但审稿人对于文章仍有如下疑惑，希望作者予以解答。

作者回应：谢谢专家的肯定。

1.Chiu 2009 提出采用聚类方法进行认知诊断时，同时采用了 k-means 和系统聚类法，而作者在此只采用了 k-means 方法。作者为何进行这种选择，希望能够给予解释。

作者回应：谢谢专家的问题。Chiu(2009)文章中提到：在 Punj and Stewart（1983）的聚类方法综述中提到，在具备先验知识的情况下，K-means 方法优于系统聚类法，同时不同距离测量方法对 K-means 方法的影响较小。Chiu(2009)模拟研究表明，K-means 的初始中心选取有两种方法，一个是将理想掌握模式作为初始中心（本文借鉴其思路），另一种是将系统聚类法得到的聚类中心值作为初始中心，结果表明，在各实验条件下，K-means 方法较系统聚类法好，尤其是当各类数据比例相差不大时。本文模拟的数据分布为均匀分布，各类数据个数几乎相当，且初始中心的选取也是前一种方法，因此较适用于 K-means 聚类法，今后的研究会进一步探讨 K-means 方法在各类数据比例不同时的表现（如正态分布和其它分布时）。

2.作者在模拟研究和实证研究中都探讨了这种新方法的知识状态分类精确性，认为新方法的表现较高。但作者并未讲新方法的表现与已有的参数认知诊断方法（如多级计分的 DINA 等）以及非参数诊断方法（如神经网络和支持向量机等）的结果进行比较。在没有比较的情况下，如何能够证明新方法的优点呢？

作者回应：感谢专家的提问，专家的意见非常中肯。是的，由于本文的主要目的是把 0-1 计分的聚类诊断法拓展到多级计分，以吻合测评实践的需要，因此研究的问题主要包括三个方面：一是多级计分聚类诊断法的算法；二是多级计分聚类诊断法判准率的影响因素；三是多级聚类诊断法在实践中的表现。由于篇幅的限制，并未直接对该方法与其它参数或非参数方法进行比较研究，因此所的结论并非直接的，而是相比其它研究，在相似条件下的间接比较结果，在下结论时未免有些太直接，我们在结果和结论部分会对措辞进行修改。但此方法的优势有一点是值得肯定的，那就是适宜于小样本数据的诊断，这是该方法的优势也是非参数方法的优势。非常感谢专家的建议。以下是对我们对作出其它间接结果的解释，请专家审阅。

祝玉芳等人(2009)关于多级评分下的属性层次方法的研究表明，在 4 种属性层次和 4 种失误率时的表现为，当失误率为 5%时，线型、收敛型、发散型、无结构型下的模式判准率 PMR（取最高值）依次为 94.2%、94.3%、95.4%、99.6%，除无结构型下的题目数不同外（其无结构型是 64 题，而本文却只有 22 题），本文模拟条件与失误率的设置方法与其相同。在题目数和样本容量均较少的情况下，本研究在 5%的失误率时，PMR 最低达到 97.2%，在 10%失误率时，最低也达到 94.6%，且最高均达到 1。涂冬波等人(2010)开发的 P-DINA 在 6 个属性的情况下，线型、收敛型、无结构型的 PMR 依次为 95.3%、94%、80.7%，且表明属性个数不宜超过 7 个，否则属性个数的增加会导致模式判准率的降低，而本研究在属性个数为 7 个、失误率为 20%及样本容量仅为 100 时，最低的 PMR 也能达到 81.5%；5%时，最低在 97.2%;10%时，最低在 94.6%。田伟等人(2012)开发的多级评分项目下的规则空间方法的模式判准率，当失误率为 5%时，线型、收敛型、发散型、无结构型下的 PMR 依次为 88.2%、84.4%、36.1%、21.1%，均低于本文的模式判准率，98.1%、97.5%、100%、99.7%（除无结构型下的题目数不同外（本文更少），本文模拟条件与失误率的设置方法与其相同）。对比前人在相似条件下的研究发现，本文采用的多级聚类诊断方法的判准率还是很高的。当然，本文只是跟心理学报发表的几篇代表性的研究结果进行了间接比较，由于篇幅限制，没有一一与其他方法进行比较，也没有把不同模型本身作为自变量进行模拟研究，我们希望能在今后的研究中进一步进行更详细的模拟设计，与其它参数与非参数方法进行直接比较研究，得到更可靠的研究结论。 谢谢专家的建议。

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.

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.

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.

3.作者与 Chiu 2009 的研究相比，作者只是把 0 1 计分拓展到了多级计分，并且在模拟研究中加入了属性层级结构和被试失误概率这两个变量。使用属性合分的方法，从二级计分拓展到多级计分是一个并不大的改动，让审稿人不禁思考本文创新性是否足够。此外，作者并未阐述为何要选择属性层级结构和被试失误概率作为变量，是否这两个变量与多级计分的特点是否有什么特别的联系。

作者回应：谢谢专家的问题。首先，关于创新性方面，本文的研究目的很简单，就是希望把 01 计分这种简便易行的聚类诊断方法推广到多级计分，以吻合测评的实践需要。因为按照非参数方法的特点，它应该是比较简单、灵活且不太受样本容量影响的。由此，本文的研究问题就是三个：一是多级计分聚类诊断法的算法；二是多级计分聚类诊断法判准率的影响因素；三是多级聚类诊断法在实践中的表现。这是此类研究的一般化思路(如田伟, & 辛涛，2012；祝玉芳, & 丁树良，2009；涂冬波，蔡艳，戴海琦, & 丁树良，2010)，由于聚类方法本身比较简单，所以本文从 01 计分推广到多级计分时，其算法当然就相较参数方法容易，从而好像显得本文的创新性不够，但就提出一个方法及这个方法的适用性到底如何而言，本文确能独成一个完整研究，且其目的性和问题性还是很明确的。当然，多级聚类方法还有没有其他的算法思路，及属性不等权重时其思路如何，这是未来要考虑的问题，也请专家进一步指点迷津。

其次，关于为何选用属性层次结构及被试失误率作为研究变量。在认知诊断评估中，影响判准率的因素很多，但其中最重要的两点是有效的认知诊断测验和适宜的认知诊断模型(Borsboom, Mellenbergh, & van Heerden, 2004)。本文探讨的是一种方法，当然就只有从测验方面寻找原因。**在认知诊断测验方面又有两个方面因素，一是被试因素，一是测验本身的因素。就被试而言**，在实际应用中，被试的作答都存在失误，失误的大小是不可控的，本文采用失误率作为变量，是想看该方法的判准率是否会因失误率的增加而急速下降，若是如此，则该方法太不稳定，是有缺陷的，这也是为什么多数探讨方法的研究都要考虑失误率这个变量的原因所在(田伟, & 辛涛，2012；祝玉芳, & 丁树良，2009；涂冬波, 蔡艳, 戴海琦, & 丁树良，2010)。在本研究中，模拟研究表明，即使失误率加大，判准率的下降也是在可承受的范围之内，表明该方法是较为稳定的。

此外，**就测验而言**，由于认知诊断评估的步骤是先确定属性及其层级关系，然后确定 R 矩阵，由此推导简化 Q 阵，基于简化 Q 阵得到认知诊断测验，从而进行认知诊断评估。在此过程中，Q 矩阵是认知诊断测验编制的蓝图，Q 矩阵是否适宜不仅关系着认知诊断测验的质量，也关系着认知诊断评估的准确性(Henson & Douglas , 2005; Cheng, 2009; DeCarlo , 2010; de la Torre, 2008;Rupp &Templin,2008;丁树良，毛萌萌等人，2012；丁树良，汪文义等人，2011)。而简化 Q 阵来自于 R 矩阵，R 矩阵表达的就是属性层级关系，层级关系表达的是属性之间的内在逻辑关系，这种逻辑关系的不同，即属性之间关系的紧密程度不一样，会不会影响判准率，这也是许多研究试图探索清楚的问题。以往研究 (田伟, & 辛涛，2012；祝玉芳, & 丁树良，2009；涂冬波, 蔡艳, 戴海琦, & 丁树良，2010) 表明，属性之间的紧密程度会影响判准率，逻辑关系越密，则判准率越高。则对于本研究所提出的非参数诊断方法，是否也是呈现这样的规律呢，这是本研究之所以选择层级关系作为变量的动因。模拟研究表明，在属性层级与判准率的关系规律上，本研究确实得到了与以往不同的结论。

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.

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.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological review, 111*(4), 1061.

Cheng,Y,(2009).When cognitive diagnosis meets computerized adaptive testing: CD-CAT.Psychometrika,74(4),619-632.

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.

DeCarlo,L.T.(2011). On the analysis of fraction subtraction data: The DINA model, Classification, latent class sizes, and Q-matiox. Applied psychological Measurement,35(1),8-24.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191-210.

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.

第二位审稿专家意见以及相关的回应

1.在 2.4 描述聚类过程时，对聚类方法实现认知诊断有一些疑问：按照 K-mean 聚类算法，一般聚类结束后所得到的聚类中心是与原始聚类中心不同的，整个聚类过程某种意义上可以说就是迭代计算恰当的聚类中心的过程。而在本研究中将 IMP 对应的能力向量作为初始聚类，这个含义可以理解，与原有的认知诊断思想吻合。但是经过聚类迭代过程，最后所得到的稳定的聚类中心，按道理来说应该是与原始聚类中心不同的，如果不同的话该如何解释，这时的聚类中心如何与 IMP 联系呢？可否进一步解释下聚类结束后，将每个被试能力向量分到某一类中，如何根据这一类来求取被试知识状态也就是 IMP 的呢？

作者回应：谢谢专家非常中肯的问题。对于这个问题，我们是这样认为的，请专家审阅。首先，在认知诊断评估中，判准率的影响因素众多，其中属性层次结构是否正确、Q 矩阵是否包含 R 矩阵是一个非常重要的前提条件，即 Q 矩阵包含 R 矩阵才能实现理想反应模式与理想知识状态的一一对应(丁树良，毛萌萌，汪文义，罗芬, & Ying, C. ,2012；丁树良，汪文义，& 杨淑群,2011)，这种一一对应才不至于几个知识状态对应同一个理想反应模式的混乱局面，从而才能提高模型的判准率。在本研究中，模拟研究的前提假设是属性层次结构是正确的，并且 Q 矩阵直接是由 R 矩阵推导得出，因此符合属性层次结构的知识状态的个数是确定的，且理想反应模式与知识状态的关系也是一一对应的。在 K-means 算法中，以各知识状态下的能力向量为初始聚类中心，具有某种知识状态的被试在理想作答的情况下，得到的能力向量与该知识状态对应的初始中心的距离为 0。在模拟被试观察反应模式时，总是要让其发生失误的，若被试发生失误作答，则会偏离初始中心，此时得到的能力向量与该知识状态对应的初始中心距离不为 0。然而，由于 k-means 聚类是根据被试的能力向量将被试分配到最近的聚类中心，每个被分配到该聚类中心的被试与该聚类中心的距离都是最近的，则即使重新计算聚类中心，只要此时的聚类中心变化不大，相对较为稳定，则被试理应还是被判归到该聚类中心，因为是采用模式与模式的距离最小，因此稳定性还是较大的。当然，也有可能会变动，这也就是为什么在被试失误较多时，各类诊断方法的判准率都不会是 100%的原因所在。在下图（图 1 和图 2），以收敛型、样本容量 100、失误率 5%为例，我们列出了其初始聚类中心（图 1）和最终聚类中心（图 2）的对比，可以发现其中心变化是不大的，相对还是较稳定的，因此能保证本模拟研究结果的正确性。以上是模拟研究情境下的解释，虽然最后中心是变化的，但是如果是以某种知识状态来解释该类被试的话，也是最接近初始聚类中心对应的知识状态的。除非还有其他的知识状态，出现此情况，则说明属性层次结构本身就是错误的，这在模拟研究中是不会发生的。其次，在实证研究中，情况会变得比较复杂，因为被试的失误是随机发生的，且失误的情境是不固定的，在加上实证研究中，R 矩阵并不一定完全正确，这样即使 Q 矩阵包含了 R 矩阵，也并不一定能保证理想反应模式与知识状态的一一对应，因此这时候极有可能会出现被试指向几种知识状态的情况，这也是为什么现在认知诊断不能百分百精确诊断的原因之一，也是实证研究中并没办法考证到底哪一种方法更好的原因之一。然而，专家的意见倒是给了我们一个继续研究的方向，其一是当初始聚类中心与最终聚类中心变化很大时，采用何种方法如何定一个切点来重新对被试进行判归，或判断这种变化是否是 Q 矩阵错误带来的。其二是对 Q 矩阵正确和 Q 矩阵错误情况下的聚类中心变化的结果进行比较，以进一步完善聚类诊断分析法的理论研究。目前，第二种研究正在设计当中。再次感谢专家的宝贵意见。

收敛型						
0	0	0	0	0	0	0
1	0	0	0	0	0	0
1	1	0	0	0	0	0
1	1	0.6	0.4	0	0	0
1	1	0.4	0.6	0	0	0
1	1	1	1	0	0	0
1	1	1	1	1	0	0
1	1	1	1	1	1	0
1	1	1	1	1	1	1

图 1 收敛型的初始聚类中心

100, 5%						
0.019231	0	0	0	0	0	0
0.94318	0	0	0	0	0	0
1	1	0.028571	0.028571	0	0	0
0.97917	1	0.63333	0.43333	0	0	0
1	0.98095	0.39333	0.56667	0	0	0
0.98438	1	0.95625	0.95625	0	0	0
0.98864	0.98701	1	1	0.9697	0.090909	0
1	1	0.99091	0.99091	0.93939	1	0
0.9875	1	1	1	1	0.9	1

图 2 样本容量 100、失误率 5%时收敛型的最终聚类中心

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.

2.在结论部分，作者提出在各属性层次结构下，其判准率均可与参数模型相媲美，特别是在发散型和无结构型下，也能达到很高水平。作者在研究展望部分也提到“本研究结果表明该方法在发散型和无结构型时也表现出了很高的判准率，其可能原因是发散型和无结构型下各属性被测量次数较其它结构多。”既然有可能是这种原因造成的，那所下结论就不够充分。可以看到在研究数据中，线型和收敛型结构考察的题目数分别为 7 个，8 个，而发散型和无结构型所考察的属性个数为 25 个，22 个，明显这两种属性结构和前面两种属性结构所考察的题目数量并不均衡，在没有排除这一变量的影响下就下结论未免欠妥当。

作者回应：谢谢专家的意见。关于属性层级结构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，并且，本文采用能力向量作为

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。

当然，层次结构对判准率的影响，有可能与题目数不均衡有关，但前人并未作此猜测与探讨，只因本文所提出的非参数诊断方法所得结论与前人研究稍有不同，为谨慎起见，才作此猜测。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.

3.结论部分作者提到“在各属性层次结构下，其判准率均可与参数模型相媲美”实际上，在这篇文章的研究设计中，并没有将非参数的聚类方法直接与参数方法的诊断效果对比。只是单纯的在非参数情境下展示判准率指标，实际上判准率指标的影响因素有很多，在其他因素都控制的情况下，直接比较参数方法和非参数的方法的判准率，再下结论才更有说服力。

作者回应：感谢专家的指点。是的，本文并未对各种方法的效果进行直接的比较，而是对比相似条件下的结果，得到的间接结论。关于这点的说明，在第一位专家的第二个问题中已经进行了详细解释，这里不再赘述。此外，我们已经在文中斟酌用词，进行了修改。并且，我们也希望在后续研究中进一步与其他方法进行直接比较，得到更可靠更有说服力的结论。

4.在 4.2.2 部分模拟被试的观察反应模式时写道“假设每个被试在每道题目上的作答失误率为 10%，先产生一个服从均匀分布 U(0，1)的随机数 r，若 r>0.95 且该题得分不是满分”这句话是不是有问题啊？在前面研究设计部分，作者说模拟三种不同失误水平的数据，这里也并没有充分说明。

作者回应：感谢专家指出该问题。可能是本文表述过于简略，以致不易理解。本文关于观察反应模式的模拟是参照已有研究(祝玉芳，丁树良,2009;田伟，辛涛,2012，以及其它类似研究)，采用以往惯用的方式模拟的，即观察反应模式是由期望反应模式的分量加上随机误差所得。具体而言，以模拟发生 10%失误率的观察反应模式为例，在期望反应模式的基础上，先产生一个服从均匀分布 U(0，1)的随机数 r。当 r<0.05、r>0.95、0.025≦r≦0.975 时，观察反应模式按如下规则获得：如果 r>0.95 且期望反应模式的项目得分不是满分，则该项目得分增加 1 分；当期望反应模式的项目得分是满分时，则该项目得分减 1 分；如果 r<0.05 且期望反应模式的项目的得分不为 0 分，则该项目得分减 1 分；如果期望反应模式的项目得分为 0 分时，该项目得分增加 1 分；如果 0.05≦r≦0.95 时，则期望反应模式的项目得分不变。由此，通过改变期望反应模式的原有分数，在随机的 10%的项目上发生失误，从而得到具有随机失误的观察反应模式。关于此部分内容，我们已经在文中进行了重新表述，请专家审阅。

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

感谢专家给出的宝贵意见，我们尽量遵照专家的意见进行修改。在意见回应部分，对专家的意见我们用红色标注，我们的回应是黑色字体；在文中，结合专家意见，我们增加、重写的地方，也用红色标注，以便专家审查。以下是对专家意见的回应。

第一位审稿专家意见以及相关的回应

感谢作者对于诸多问题的详细解释和回答。

作者在对比新方法和之前若干方法被试分类效果时，坚持只与前人设计相似的研究进行比较，而不愿在自己研究中进行条件一致的模拟比较。我的建议是能不能把相似条件下的结果以表格的形式列出，这样便于理解和比较。文字说明实在是有些抽象，并且条件较多容易弄混。希望作者能够再思考一下结果对比的呈现方式，使得自己的结果更有说服力。

作者回应：谢谢专家对于我们工作的肯定。鉴于专家再次指出要看看模拟条件一致时，本方法与前人方法的直接比较结果，听从专家建议，为增强本研究结果的说服力，**在本文讨论的第一部分**，加入与前人研究完全相同条件下（多级计分方法）的比较（田伟和祝玉芳等人对等级反应模型的规则空间方法和 AHM 方法（包括 A 方法、B 方法、LL 方法））。考虑到篇幅和时间限制，挑选线型条件来做比较（因在本研究中，线型结构下的判准率较其他结构低，而前人研究是线型条件下判准率更高）。

与前人实验条件、被试的理想反应模式以及观察反应模式的模拟方法全部相同，实验条件为：被试总分服从正态分布、被试人数为 5000 人、属性层次结构为线型（7 题）、失误率分别

为（2%、5%、10%、15%），各实验条件重复 30 次，结果如下（表中数据均为 30 次的均值）：

表 1 多级聚类诊断法与前人方法的比较

方法	分析水平	2%	5%	10%	15%
多级聚类诊断法	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

从表 1 可以看出，在与前人模拟条件完全相同的情况下，该方法的判准率表现出一定的优势，尤其是在失误率增大的情况下，该方法的模式判准率表现出了更好的稳定性。

第二位审稿专家意见以及相关的回应

关于聚类中心点变化的问题，审稿人有以下思考：
聚类分析是一种探索性统计分析方法，其与判别分析不同，判别分析是已知研究对象用某种方法分成若干类的情况下，确定新的观测数据属于哪一类。而在本研究中所采用的聚类分析确定初试值的方法，若聚类中心点变化不大，更像是判别分析。所以这种方法的分类思想与规则空间模型很接近。那么这种方法的独特之处在哪里呢？虽然计算简洁，但是其实k-mean聚类采用的欧氏距离的计算方法很不稳定，并且初始值的确定对聚类结果影响很大，这些问题在文章Chiu(2009)中，原创作者已经指出。作为在这篇文章基础上的研究，作者虽然将该方法拓展到了多级计分模型，但是原有文章中抛出的一些问题并没有得到深入的解决。有重复研究的嫌疑。

作者回应：感谢专家的深刻剖析。首先，我们赞同专家关于聚类分析和判别分析区别的分析。为什么本文用“聚类分析”而不用“判别分析”这个词，我们是这样认为的：判别分析是已知每个被试从属于某个类，以此分类变量为因变量，以分析指标为自变量，建立判别方法，用此方程可判别未来未知被试从属于哪一类，判别分析的主要目的在于预测。而在本研究中，尽管我们基于 Q 矩阵可算出被试可分为多少类（理想掌握模式数或理想反应模式数），但并不知道哪些被试到底属于哪一类，而需要把被试的观察反应得分与理想反应得分一一比较后归为距离最近的类。再者，本文是受 Chiu(2009)的启发而做的继续研究，Chiu(2009)用的是“聚类分析”这个词。基于以上两点，本文还是采用“聚类分析”这一说法。

其次，专家认为此方法与规则空间法分析思想相同，确实如此。这两种方法都可采用计算距离的方法，但具体过程却不一样。RSM 包括建构规则空间和判别归类两个阶段，第一阶段是基于 Q 矩阵得到理想反应模式和理想掌握模式，然后利用 IRT 估计出理想反应模式被试的能力值θ和异常反应程度的警戒指标ς，(θ,ζ)构成了规则空间中的序偶或纯规则点。第二阶段根据被试在测验上的作答反应向量 X 估出实际序偶(θ,ζ)，计算该序偶到各纯规则点的马氏距离 D 或计算实际反应模式属于每个理想反应模式的后验概率，将被试分类到距离最近或后验概率最大的知识结构中。可见，RSM 需要根据 **IRT 的参数方法**算出理想反应模式和被试观察反应模式所对应的序偶(θ,ζ)，然后计算每个实际(θ,ζ)到各理想(θ,ζ)的马氏距离或后验概率，把被试归类到距离最小或后验概率最大的理想反应模式中。在本研究中，聚类诊断法是基于 Q 矩阵或 R 矩阵，算出 IMP 和 IRP，然后根据属性合分及其标准化思想得到 IRP 和被试 ORP 在各个属性上的能力向量；以 IMP 所对应的能力向量为初始聚类中心，计算各个 ORP 各属性能力向量与 IMP 各属性能力向量上的距离，如此循环，把各 ORP 归类到距离最近的 IRP 或 IMP 中。RSM 和聚类诊断法的思路见下图 1 和图 2。由此可见，RSM 通过使用 IRT 参数方法将理想反应模式和观察反应模式降维成二维规则空间，通过实际序偶到各理想序偶（纯规则点）的距离或后验概率进行被试掌握模式的归类，而多级聚类诊断法是基于属性合分及其标准化的思路，计算出理想反应模式和观察反应模式所对应的属性能力向量，通过观察反应模式的属性能力向量到各理想反应模式的属性能力向量的距离，实现对被试知识状态的归类。RSM 和多级聚类诊断法尽管都采用了距离的方法，但其思想和过程并不一样，RSM 至关重要的第一步是 IRT 参数方法，需要样本容量较大，而多级聚类诊断法只是运用到总分和标准化思想，并未用到参数方法，因此可适用于小样本情境，且 RSM 是通过序偶的比对进行判别，聚类法是直接通过属性得分的比对进行判归。由此，聚类法不仅计算简便且更适宜小样本数据，并且判准率比 RSM 高出很多（见对第一位专家的回应）。所以，我们认为该方法与 RSM 还是有区别的，是关于距离分析的另一种思考。

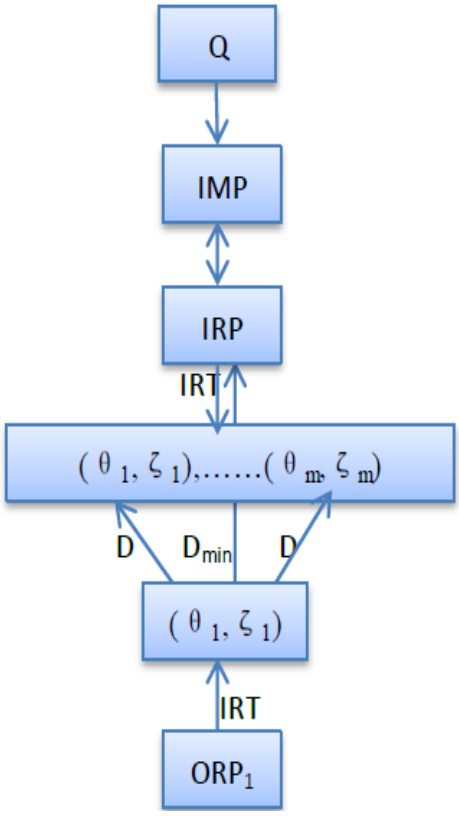

图 1 RSM 认知诊断思路图

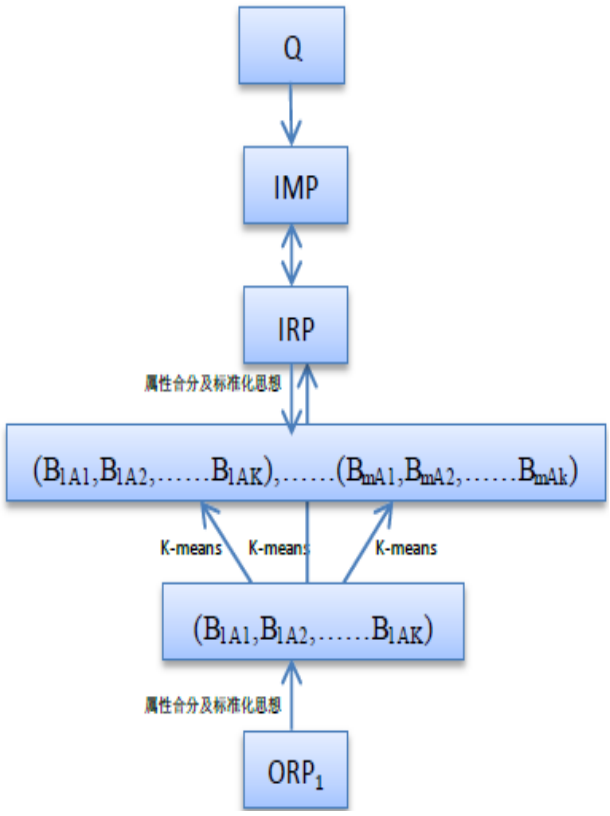

图 2 RSM 认知诊断思路

最后，关于本研究的目的、内容与Chiu（2009）重复研究嫌疑的问题，我们拟对两篇文章的内容进行比对和说明，见下图3和图4。**Chiu(2009)的研究内容为**：1.前言部分，提出多元统计分析中的内容如聚类分析、判别分析和多维度量尺分析等可以实现对被试的认知诊断与分类，并且这些方法具有操作简单，无需参数估计等优点。2.认知诊断潜类别模型部分，主要介绍了几种认知诊断方法或模型，重点介绍了DINA模型，因为该文的数据模型是基于DINA模型的。3.认知诊断中的聚类分析部分，先是介绍了属性合分方法(sum-scores)，然后介绍了聚类分析中的各种距离测量法，其中重点介绍了系统聚类法(HACA)和K-means方法，因为该文的主要目的之一是比较这两种方法的优劣。4.分类的理论证明部分，主要用数理的方法对聚类分析可用于认知诊断分析进行了理论证明，篇幅较长，是本文的主要部分也是其主要贡献。5.模拟研究部分，也是本文的重点，主要目的是比较DINA、HACA和K-means三种方法在诊断分类中的一致性，及对后两种方法进行比较。模拟数据是基于DINA生成的，主要条件有三个：属性K（3个和4个）；样本容量N（100和500）；题目J（20、40和80）。以ARI(Adjusted Rand Index)和 ω 类内同质性两个指标作为结果变量。6.语言数据的实证研究部分，对R-RUM、HACA、K-means在语言数据上的表现进行了比较。7.讨论部分，对聚类分析的优劣进行了讨论。指出尽管该方法还有许多问题需要进一步研究，但其在认知诊断分类中的作用和优势是显而易见的。**本论文的研究内容为：**1.前言部分，在肯定参数模型诸多优势的同时，指出其局限性，而非参数方法可从另一个角度进行诊断分类，解决计算繁杂适合小样本课堂评估等问题。接着在阐述非参数方法特别是属性合分思想（Henson，2007）、Chiu(2009，2013)聚类分析、 Ayers等人对属性合分标准化（2008，2009）思路的基础上，结合当前考试改革的现状，指出本文的研究目的：本文拟将聚类诊断分析这种简单易行的非

参数诊断方法拓展至多级评分，同时探讨不同样本容量、不同失误率及不同属性层次结构下，该方法的诊断正确率。2.0-1计分聚类诊断方法的简介，受Chiu(2009)及前言中提到的相关文献的启发，系统阐述了0-1计分聚类诊断方法的思路，并以具体实例详细阐明了整个过程。这在Chiu(2009)中只是点到为止的，而我们通过自己的理解把它操作化具体化了。尤其是当属性个数不同时，我们借鉴了标准化的思想，以消除属性数目带来的影响。3.把0-1计分的聚类诊断法推广到多级计分。这一部分，多级计分的属性合分及其标准化思路是作者首次提出的，并以具体实例说明了整个过程。4.模拟研究部分。为验证多级聚类诊断法的效果，采用模拟的思路，探讨属性层次结构，样本容量、失误率对判准率的影响，以考察多级聚类诊断法的稳定性及适用性，并对其结果进行了详细分析。5.实证研究部分。考察多级聚类诊断法在数学应用题认知诊断中的适用性，并与多级规则空间的结果进行了简单比较。6.讨论部分。结合本研究目的与结果，探讨了该方法的优势与本研究的发现。7.结论部分。对本研究的发现进行总结，对其不足和可能存在的问题进行了阐述，并指出了未来研究展望。

由上可知，这两篇文章的目的和重要内容都存在比较大的不同。Chiu(2009) 主要重点在聚类法适用性的数理证明及不同聚类方法或距离测量法的一致性及与 DINA 模型的比较方法。并且，由于 HACA 只能将被试聚类，使组间差距大，组内差距小，不对各类标定，因此无法确定被试属于哪类，同时可能分的类数也与理想掌握模式的个数不同。因此要比较 K-Means 和 HACA，只能比较两方法的同类一致性。但 K-Means 本身是可以对被试进行标定的，即比较被试与最后得到的所有聚类中心的距离，距离最近那类为被试的分类结果。Chui 也认为在多数实验条件下，K-means 聚类的结果更好（故本研究用该方法）。所以，不可否认，Chiu（2009）的研究为聚类分析法在认知诊断中的应用提供了思路和数理证明，是其重大贡献。我们的研究正是深受其启发才有思路的。然而，我们的研究目的与重点是与其不同的。首先，Chiu(2009)指出属性合分思想在聚类中的应用，但并未考虑属性数目不同带来的差异，未采用标准化思想，而使用欧式距离时，各维数据最好单位相同，本文使用的能力向量是对合分向量进行标准化后的结果；其次，本文将该方法扩展为适用于多级计分数据，提出了多级计分情境属性合分及其标准化的思路，并以实例重点阐述其过程，使本方法具有可操作性；再次，在模拟研究中，分析了不同属性层次结构和不同被试人数及不同失误率对该方法的影响，以考察该方法的适用性及稳定性，并对结果进行了详细分析，发现了一些有意义的结论（见文中）。需要说明的是，本文的模拟思路、条件及评价指标与 Chiu(2009) 都是不一样的，Chiu(2009) 由于是基于 DINA 模型的（结果易受 EM 算法的影响，而本研究模拟数据的生成并不依赖任何模型），不考虑层级结构，所以其基础 Q 矩阵是 20 题时的 3 个和 4 个属性时的 Q 矩阵，而 40 和 80 题是在 20 题的基础上扩大 2 倍和 4 倍。如果用丁树良老师的研究来说，Q 矩阵包含的 R 矩阵越多则判准率越高的话，假如 20 题时的 Q 矩阵包含了一个 R 矩阵，则 40 和 80 题时的 Q 矩阵包含的 R 矩阵个数是 20 题时的 2 倍和 4 倍，则 40 题和 80 题时与 20 题时的结果差异的归因可能并不是题目数多少带来的，而是含 R 矩阵个数多少带来的。Chiu(2009)并未意识到这点。

因此，综上所述，我们认为本研究与 Chiu(2009)并不存在重复研究的嫌疑，请专家审阅！

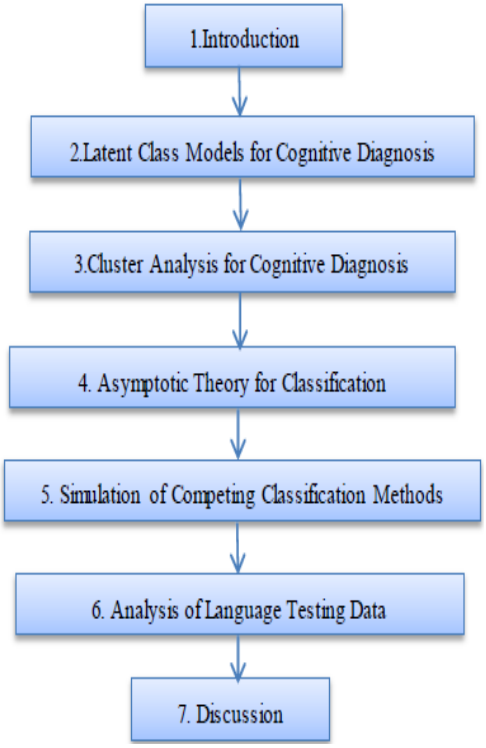

图 3 Chiu(2009)内容框架

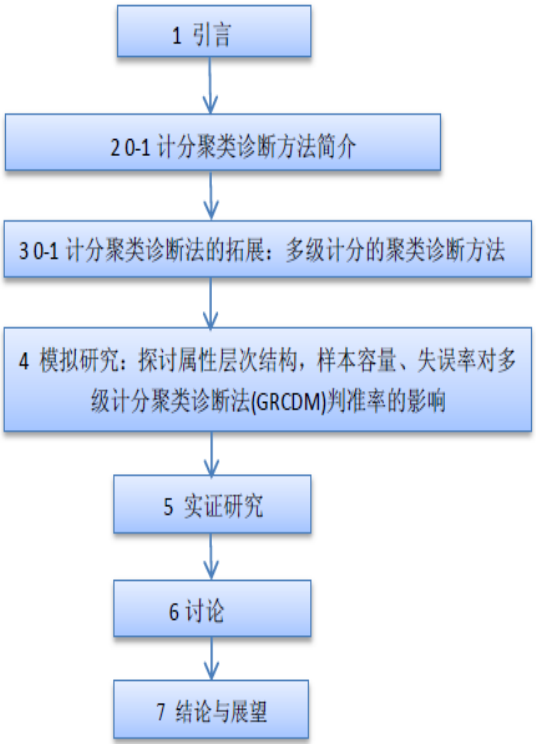

图 4 本研究内容框架

在作者的回复中所提供的数据，以失误率为5%这种情况下所提供的数据，聚类中心点虽然变化不大，但是已经发生变化了，那么随着失误率的增加，聚类中心点一定也会跟着变化的，这种变化的幅度如何呢？

作者回应：谢谢专家。是的，随失误增大，中心变化的幅度会增大，这也就是为什么随着失误率增大，判准率也会下降的原因所在，然而无论是聚类中心的变化幅度还是判准率的变化幅度，都在可接受的范围之内（以线型为例，因本研究中，线型的判准率最低）。表 2 为最后得到的各类中心值与其初始中心值之间的距离（变化幅度）。采用欧式距离，不同的初始中心之间的欧式距离最小（最近的两类，即只有一个属性不同，如（0001101）与（0001111））是 1。表 2 说明聚类中心的改变比起初始聚类中心之间的距离小很多。同时还可以看出，随着失误率的增大，的确会导致最后聚类中心与初始中心发生偏移增大，其实从模式判准率和属性边际判准率的下降也可以看出，然而其变化幅度和判准率都在可接受的范围之内。而如果聚类中心本身变化很大，极有可能证明本身的层次关系或者 Q 矩阵就有错误，至于如何修正，以及变化多少说明 Q 矩阵发生错误，这个问题需要继续深入研究，本研究还未涉及。

表2 线型结构下聚类中心的变化幅度

失误率	第 1 中心	第 2 中心	第 3 中心	第 4 中心	第 5 中心	第 6 中心	第 7 中心	第 8 中心	平均
5%	0.038961	0.063917	0.043643	0.06717	0.056142	0.08916	0.046291	0.036346	0.0552
10%	0.085714	0.11624	0.093449	0.11138	0.091225	0.10788	0.064007	0.062163	0.0915
20%	0.17922	0.19213	0.15311	0.18672	0.14498	0.18289	0.11148	0.13415	0.1606

作者也提到深入研究的方向：对Q矩阵正确和Q矩阵错误情况下的聚类中心变化的结果进行

比较。审稿人的理解为：即使在Q矩阵正确的情况下，现在还不能确定聚类中心点的变化是否会影响诊断效果。

作者回应：谢谢专家。若Q矩阵正确，聚类中心的变化随失误率的增大是会呈现变大趋势（见表2），但是在失误率为20%时，聚类中心变化也是在可承受的范围，比1小很多，且判准率也并不低，PMR最低也能达到.815(样本容量为100时)，最高达.978，同样的情形可见上表1。然而，尽管如此，本方法肯定还有需要进一步发展和深入研究的地方，因为认知诊断发展至今，并没有完美的模型（方法），每种模型都有其适用条件、前提假设、优势与缺陷，这也就是一种方法的提出，就会有很多学者从不同的角度进行不同的研究解决不同的问题，使研究不段的推进。毫无例外，本方法也存在需进一步发展和完善的地方。我们定会继续前行！

另外一个思考就是传统的参数认知诊断方法虽然计算过程较繁琐，但是它是一个独立的专用于认知诊断的方法，其中的参数水平某种程度上已经表征了所采用的Q矩阵的合理程度，通过模型与参数的拟合程度也能根据具体不同的数据集选择合适的模型。这些方法在聚类诊断方法中如何解决呢？

作者回应：我们认同专家关于参数方法的看法。本研究并没有否定参数方法的优势，相反，目前多数研究都是参数模型，经过近20年的发展，其研究也较深入，近年来出现了拟合的方法。然而，参数方法和非参数方法本身就各有优势与缺点的。参数方法有其优势的同时，也有局限性，往往需要大样本数据，参数估计过程复杂，对于课堂评测及非专业人员确实是有些无奈，因此受Chiu(2009)的启发，本文对非参数方法提出了自己粗浅的研究，然而，非参数方法因其没有参数，它对数据处理的思路是与非参数方法不同的，至于模型选择与拟合检验的问题，还需时间和更多的研究者来进行，本文仅起到抛砖引玉的作用。

关于本文的研究设计方面，审稿人有以下建议：作者并没有将聚类方法的诊断结果与参数模型的认知诊断方法直接比较。虽然作者在文章中与其他研究进行了间接比较，但是每一个研究都有独特的假设情境，与本研究的情境并不完全对应。所以建议作者在本研究设计中加入直接的比较结果，方才更有说服力。

作者回应：谢谢专家建议。鉴于与前人研究相似并不足以说明该方法的优势，因此，**在讨论的第一部分**，加入与前人研究完全相同条件下（多级计分方法）的比较（田伟和祝玉芳等人对基于等级反应模型的规则空间方法和AHM方法（包括方法A、方法B、方法LL三种方法）的研究）。考虑到篇幅和时间限制，挑选线型条件来做比较（因在本研究中，线型结构下诊断率正确相较其他结构较低，而前人研究是线型条件下判准率更高）。

与前人实验条件、被试的理想反应模式以及观察反应模式的模拟方法全部相同，实验条件为：被试总分服从正态分布、被试人数为 5000 人、属性层次结构为线型（7 题）、失误率分别为（2%、5%、10%、15%），各实验条件重复 30 次，结果如下（表中数据均为 30 次的均值）：

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

从表中可以看出，在与前人模拟条件完全相同的情况下，该方法的判准率表现出一定的优势，尤其是在失误率增大的情况下，该方法的模式判准率表现出了更好的稳定性。

以上是我们的第二次修改说明，再次感谢专家的辛苦审阅，并祝专家新年快乐！

第三轮

感谢专家给出的宝贵意见，我们尽量遵照专家的意见进行修改。在意见回应部分，对专家的意见我们用红色标注，我们的回应是黑色字体；在文中，结合专家意见，我们增加、重写的地方，也用红色标注，以便专家审查。以下是对专家意见的回应。

感谢作者对所有问题的思考和耐心解答。
认知诊断现在有很多理论研究，但是这些理论研究的前提假设是在 Q 矩阵构建恰当的前提下，如何从根本上推进 Q 矩阵的构建以及验证，或者推进认知诊断理论的应用研究更是当务之急。否则在此基础上很多变换的方法投入实际应用的可能性有多大呢？

作者回应：感谢审稿专家的意见。关于这些问题，我们是这样看的：

首先，关于理论研究的问题。正如专家所言，测量理论发展的任何阶段，都可分为理论研究和实践应用研究两个取向，并且往往都是理论研究先向，实践应用滞后。这是符合事物发展规律的，因为理论必须先行才能指导实践，但同时理论也必须跟实践相结合。在认知诊断的理论研究中，主要包括对模型开发、模型融合、参数估计、Q 矩阵的充分性及修正等。这些研究多数从模拟研究的角度出发，先假定 Q 矩阵正确的情况下，然后通过模拟实际情境中的各种情况，如被试失误、猜测，Q 矩阵误设等，尽量产生吻合实践情境的数据，然后通过各种条件下的模式判准率、属性判准率等一系列返真性指标，来对先前提出的假设、方法、模型进行验证。因而，相比实践研究，模拟研究在 Q 矩阵的构建方面更灵活，可以假定其正确或者误设等，但是，我们认为理论研究中采用的模拟设计，并不是任意设计的，而是尽量考虑实践情境，跟实践情境吻合的设计，因而其研究成果是对实践研究有用的，并且

可以指导实践研究，如结合各种模型的前提假设条件、优势与不足，选择适宜的模型进行实践研究中的数据分析；再如更加 Q 矩阵充分性的原则，在实践的诊断测验编制时尽量保持属性与项目的平衡、Q 矩阵包含一个或多个 R 矩阵等。

其次，关于应用研究的问题。诚如专家所说，在实践应用研究中，Q 矩阵构建非常重要，因为它是测验编制的蓝图，Q 矩阵是否完备和充分，关系着认知诊断测验的质量，从而关系着认知诊断分类的判准率，直至最后关系着对每个学生认知结构完善的补偿教学。因此，如何构建完备的 Q 矩阵，以及如何对 Q 矩阵进行验证也是目前关注的一个热点问题。关于这项研究，我们也在努力研究和工作中。比如，我们单位的基础教育质量监测研究中心目前就主建了“科学、数学”两个学科的专家队伍，对中学科学和小学数学相关领域的认知模型进行构建，尤其是在小学数学“分数的意义和性质”、“行程问题”、“四年级应用题”、“六年级图形与几何”等几个领域。通过跟一线教师的共同努力，我们已经形成了初步的研究成果。并且，关于这些领域 Q 矩阵的构建及修正过程，Q 矩阵完备性和充分性的验证等相关成果会另外撰文表述。我们想说的是，关于应用研究，我们也一直在前行中。

再次，关于投入实际应用的问题。是的，测量理论发展至今，都存在理论研究超前而实际应用滞后的问题。然而，我们可以这样来理解这个问题，只要理论研究做的扎实又实用，随着社会的发展，这些研究终归是有用武之地的。正如心理测量专家余嘉元（2012）所言：“心理测量学，她是心理学皇冠上的数学明珠，经典测验理论、项目反应理论和概化理论是它的基础，在和计算机科学、认知科学的结合中产生的计算机化自适应测验和**认知诊断测验正在得到应用**，计算智能也开始成为研究的重要手段。**它在教育、医学、管理、工业、军事等领域都有广阔的应用前景，在建模、参数估计、等值、项目功能差异、标准设置等方面还要加强研究**，心理测量的应用也需要有良好的法制环境，以及具有高素质的人才队伍。”因此，我们认为，认知诊断理论发展至今，才短短 30 来年，理论研究不是太多，而是太少。我们都明白认知诊断评估的理念是很好的，非常吻合测评的发展性和诊断性功能，吻合现代教育关于“一个都不能落下”的思想，一线教师也能理解这个思想，但应用起来比较困难，为什么？因为背后的诊断方法太复杂，已有的方法在各有优势的同时又各有缺陷，比如参数方法太难，需要样本量太大等，限制了它的实践应用。正因如此，我们才大胆扩展本文的非参数方法，通过模拟和实证研究证明其性能，而研究结果也表明其无需参数估计，计算简便，无需前提假设、不受样本容量影响等，可以适合小样本测试及课堂评估。本文的目的也正是想通过表述这个研究的结果，为推动认知诊断评估的实践应用略尽绵薄之力。

最后，关于我们的感想。我想，只要有一批既热爱心理测量学的理论研究，又关注其实践应用研究的人才加入到测量学的队伍中来，心理测量学必定能迎接更多的挑战与机遇，拥有更广阔的应用天空，在心理学科学化的道路上，这颗明珠定能熠熠生辉！

余嘉元.(2012). 心理测量学：心理学皇冠上的数学明珠. *中国科学院院刊, 27*:209-215.

再次感谢专家的意见，也向即将付出辛劳的复审编委致敬！

第四轮

对编委意见的回应

感谢编委给出的宝贵意见，我们尽量遵照编委的意见进行修改。在意见回应部分，编委的意见我们用红色标注，我们的回应是黑色字体；在文中，结合编委意见，我们增加、重写的

地方，也用红色标注，以便编委审查。以下是对编委意见的回应。

As there are 2 reviewers recommend publication, so I would recommend this article for publication.

作者回应：感谢审稿专家和编委客观公正的建议。

CDA should be Cognitive Diagnostic Assessment ??? please check related terms in the whole article

作者回应：谢谢编委的指正。是的，CDA 是 Cognitive Diagnostic Assessment 的缩写，我们已经对文中所有应表述为 Diagnostic 的地方进行了修改，并用红色进行了标注。

Now counting the content, not including references and tables, there are 11597 words, I would recommend cutting 2500 words to 9000 words.

作者回应：感谢编委的建议。确实，因为本文既有模拟研究又有实证研究，再加上按照审稿专家的建议，我们增加了与前人研究的比较，文章字数稍有超出。按照编委的建议，结合“投稿指南”的要求：建议来稿正文（不包括图表、摘要、参考文献）字数限制在 10000 字以内，参考文献限制在 30 条以内。我们对文章的语言进行了精炼，目前文章的字数为：不包括图表、摘要、参考文献，正文字数 9130 字，请编委审阅。如若需要，我们下次再作删减。

For the tables, are we printing them in the article? Could we remove all the Appendices, in particular, Tables 1, 2 in appendices??

作者回应：谢谢编委，我们接受编委的建议。我们之前之所以把附录中的几个表格放在文后，是想向评审专家和编委完整展现这个方法运算的整个过程，提供原始运算结果，以便专家批评指正。按照编委建议，由于删除附录中的表格并不影响本文的可读性、研究过程及结果解释，所以现在把附录中的表格统统删除了，如果有需要，我们可随时提供。

I have briefly edited the abstract for the consideration of the authors (see attached).

作者回应：非常感谢编委为本文英文摘要所进行的修改和修正工作，纠正我们的语法和用词错误。向您认真、严谨的指导工作表示敬意！