《心理学报》审稿意见与作者回应

题目：多级评分的认知诊断计算机化适应测验
作者：蔡艳，苗莹，涂冬波

第一轮

非常感谢审稿专家的细致审稿及提出的宝贵意见和建议，这些都促进了本文的进一步完善。根据专家们提出的意见及建议，我们对文章做了认真修改，现就文章修改了的部分做如下详细说明，所有修改了的部分我们均在文中用黄色进行了标注。

审稿人 1 意见：论文中方法正确，推导过程无误，表述清楚，结构合理，逻辑缜密，是一篇较好的学术论文。存在的问题在于：

意见 1：对于 EAP 方法的估计结果，形式上是否存在不足，0,1 表示的知识状态？

回应：专家的这条意见非常好。在认知诊断中，如果采用 EAP 方法估计被试的知识状态，则估计值一般是介于 0-1 之间的概率，即每个被试在每个属性上的期望后验掌握概率；而在认知诊断中，一般会根据被试在每个属性上的期望后验掌握概率划分为掌握或未掌握，而划分的依据一般是以 0.5 为标准，即当被试期望后验掌握概率小于 0.5 则断为没掌握（即 $\hat{\alpha}_{ik}=0$），否则判为掌握（即 $\hat{\alpha}_{ik}=1$）。在本研究中，被试参数估计方法采用是 MLE/MAP 法，这两种方法估计的结果则直接是 0－1 二分化的。

意见 2：进一步修改正文稿格式，错别字错误。

回应：非谢谢专家的意见，根据专家的建议我们又通读了全文，并对相关格式及错别字做了修改，详见文章第 11、13、21、22 页。

审稿人 2 意见：作者基于涂冬波等人（2010）提出的 P-DINA 模型，针对多级评分项目提出了 GP-DINA 模型，这个模型比 P-DINA 模型更合理些。作者还基于 0-1 项目的选题策略 KL、PWKL 和 HKL 给出多级评分的相应选题策略 PS-KL、PS-PWKL 和 PS-HKL。最后通过模拟研究，比较了几种选题策略的效果。该方法对多级评分 CD-CAT 的研究具有推动作用，但还有以下几个问题：

意见 1：请在文章中给出 GP-DINA 模型中猜测参数和滑动参数的限制条件。模拟中选择的参数取值在 0 至 0.6 中,是否满足了这些限制？在表 1 中，理想得分为 0,1,2 的三类被试观测得分为 3 分的概率都是 g_{j3}，理想得分为 1,2,3 的三类被试观测得分为 0 分的概率都是 s_{j1}，这样假设可能不符合实际，比如,实际情况往往是理想得分为 1 分的被试比理想得分为 3 分的被试观测得分为 0 分的概率更大。

回应：非常感谢专家的这条建议。我们在模拟时，严格控制了 s， g 从均分布$U(0,0.6)$中随机产生，并控制了$s_{jt} \le s_{jt+1}$和$g_{jt} \ge g_{jt+1}$，具体控制的方法是：第 1，如果随机产生的 s 或 g 大于 0.6 则重新产生，直至产生的值在（0，0.6）区间；第 2，对于同一题多个 s 和 g，则分别通过排序的方法，以达到$s_{jt} \le s_{jt+1}$和$g_{jt} \ge g_{jt+1}$，根据专家的建议，我们已在文中第 16 页进行了补充说明。

表 1 中，的确如专家所言，理想得分为 0,1,2 的三类被试观测得分为 3 分的概率都是 g_{j3}，这个与 DINA 模型属完全非补偿型模型（full non-compensational model）有关。在 DINA 模型中如果被试没有掌握项目测量的所有属性，则不论被试是什么其它属性掌握模式，他们答对的概率都是完全相同的猜测概率 g。例如：DINA 模型中，项目测量模式为（111），那么只有掌握模式为（111）的被试答对该项目的概率为（1－s），但其它七种类型掌握模式（如 000，100,010,001,110,101,011）的被试答对的概率均是完全相同的猜测概率 g，也就是说 DINA 模型在每题上有且仅能区分出两类被试，而每类被试的内部它是没有办法区分开的，我们也无法指望 DINA 模型在每道题上都能区分出所有的被试，这是由 DINA 模型是一个完全非补偿模型（full non-compensational model）的特点所决定了的，但这不影响 DINA 模型的诊断功能，因为随着测验长度的增加，DINA 模型会综合测验所有项目的信息，从而实现对所有被试的区分与诊断。本文开发的多级评分 DINA 模型（即 GP-DINA 模型），也基本沿用了 DINA 模型这一特征，即如果被试没有达到得 3 分理想得分的属性掌握模式，则不论被试是什么其它属性掌握模式则答得 3 分的概率也均是相同的猜测概率即 g_{j3}，这种现象也是由 DINA 模型是一个完全补偿模型这一特点决定的，但与 DINA 模型一样，这也并不影响 GP-DINA 模型实现对所有被试的区分与诊断，因为 GP-DINA 模型中：理想得分为 t 分的被试，他/她观察得分也为 t 分的概率是倾向最大的（详见表 1 阴影部分概率）；而且若理想得分为 0 分的被试，他/她猜对 1 分的概率会大于猜对 2 分的概率，猜对 2 分概率会大于猜对 3 分概率，余类推（即满足$g_{jt} \ge g_{jt+1}$）；而且若理想得分为 3 分的被试，他/她在 1 分上失误的概率会小于在 2 分上失误的概率，在 2 分上失误的概率也会小于在 3 分上失误的概率，余类推（即满足$s_{jt} \le s_{jt+1}$）等等，这些条件或限制都保证了 GP-DINA 模型能够在测验层面（不是单个项目层面）实现对所有被试的诊断与区分。而且，大量的 Monte Carlo 模拟实验也进一步表明，GP-DINA 模型的模式诊断正确率也高达 95.5%（6 个独立属性，60 题 1000 名被试），这些都表明 GP-DINA 是有能力区分出不同类型的被试。再次非常感谢专家的意见。

意见 2：文章模拟设计中提到“本题库中所有试题均采用 0-3 的四级评分方式”，而且题库中包括全部 63 种题型，从文章中看不到仅包含一个属性的项目，如何评分为 0,1,2,3 四个等级？按照文章理想得分的计算公式(4)，仅包含一个属性的项目理想得分仅有 0 分和 3 分两种可能的结果，无法取到 1 分和 2 分。包含 2 个属性的项目也有类似的问题。

回应：非常感谢专家的真知灼见，专家的这条意见非常重要。的确如此，如果题目仅测量了一个属性，那个被试在该题上的理想得分和观察得分都仅有两种（要么 0 分，要么满分），其实这时多级评分自然也就演变成了 0－1 二值评分（本文中是 0－3 二值评分），因为评分点有且仅有一个属性，因而对于只测量了一个属性项目的评分自然也就只有两种评分方式，即要么掌握了该属性得满分，要么没掌握该属性得 0 分，这也是符合我们实际评分规则的。2 个属性的项目如专家所言，也有类似情况。这在实际中也是有一定合理性的，比如作文题满分是 60 分，但并不是说 0－60 分的每个得分点都有被试，有的得分点会出现没有的情况；简答题、计算题等级评分的试题往往存在某个得分点没有的情况。因此，本研究模拟的题库中，应该既有 0－3 二级评分试题，也有 0－1－2－3 四级评分试题，还有 0－1or2－3 三级评分试题，也就是说本文模拟的题库不都是“0-3 的四级评分方式”（但满分都是 3 分），而是包含了二级、三级和四级三种记分方式，是一种混合的题库。我们后来进一步查看了模拟的题库及程序，也的确是出现了上述三种记分类型的试题（但这三种类型试题的满分均是 3 分），因此为了谨慎起见，根据专家的建议，我们修改了题库模拟的表述，详见文章第 16 页，非常感谢专家的指出。

意见 3：在 CAT 模拟中，被试回答的第 1 题是按什么原则产生的？作者没有描述，是随机选择的,还是根据猜测参数和滑动参数的值产生的?

回应：在 CAT 模拟中，我们采用的是传统 CAT 的做法，即第 1 题是按照随机的方法选的。当然，我们觉得专家提到的“根据猜测参数和滑动参数的值产生”也很有道理，这个未来可以进一步探索，谢谢专家的意见。

意见 4：在实际中，要将多级评分项目放在 CAT 中，在计算机上实现多级评分项目的快速评分问题需要考虑,比如数学考试中的综合题如何实现计算机自动评分？不知作者在这方面有什么考虑？

回应：专家提出了一个非常值得去探究的研究方向。的确，对于多级评分的 CD-CAT，项目的自动评分是非常重要的。对于心理学量表中经常用到的 Likert 型等多级评分的试题，这个完全可以由计算机实现自动评分。但诸如专家提到的数学考试中的综合题的自动评分，这个需要其它诸如计算机领域中人工智能技术的跟进与发展。比较可喜的是，英语作文评分目前国际上已实现了计算机自动评分，中文作文自动评分的技术也日益发展并不断成熟，这些技术都为本文多级评分 CD-CAT 提供了重要的支持，我们也深信随着人工智能技术以及测量技术的不断发展，多级评分综合题的自动评分将会迎刃而解。根据专家建议，我们把多级评分综合题的自动评分技术研究加入到文中最后的讨论与展望部分，详见文章第 21-22 页，谢谢专家的这条意见。

审稿人 3 意见：《多级评分的认知诊断计算机化适应测验》一文，在 P-DINA 模型基础上提出了一种多级评分的 CD-CAT 模型，这是在理论上对认知诊断计算机化适应测验的扩展，从这点上看，研究有价值。但本文的不足之处如下：

意见 1：本文是在 P-DINA 模型上的拓展，国内又有学者提出过多级评分 CAT 的认知诊断方法，在进行仿真实验时，是否应立足于比较新提出模型与原有模型相比的优势；而不是对新模型加以不同的选题策略，因为本文的研究重点并非比较选题策略；

回应：非常感谢专家的建议。关于多级评分 CD-CAT 的研究，我们查阅国内外公开发表的文献，也仅查到了一篇相关研究（见周婕等人，2007），这说明国内外对于这个领域的研究十分薄弱，亟待进一步研究。在行文时，我们也考虑了与周婕等人（2007）研究的比较，但考虑到本文研究的多级评分 CD-CAT 与周婕等人（2007）的研究从使用的计量模型（一个是多级评分认知诊断模型 GP-DINA 模型，一个是多级评分 IRT 模型加 0－1 评分规则空间模型），题库项目参数（一个是基于认知诊断的 s 和 g 参数，一个是基于 IRT 的 a 和 b 参数等），选题策略（一个是基于 CD-CAT 的 KL 类信息量选题，一个是基于图论“盖住”选题），诊断思路（一个是全程诊断，另一个是先诊断后估计能力）等方面均不同，因此两者间不具有真正意义上的可比性，因此为了谨慎起见，本研究中并没有与周婕等人（2007）研究进行比较。但就本研究而言，从方法学的角度来看，与已往研究相比，本研究最大的特点是构建了一个适合多级评分 CD-CAT 的认知诊断模型，是真正基于认知诊断的多级评分 CD-CAT，而不是采用传统的基于 IRT 方法的 CD-CAT；并且，从 Monte Carlo 模拟实验来看，本研究开发的多级评分 CD-CAT（PS-PWKL，PS-HKL 选题策略下）具有较理想的诊断正确率、曝光控制和测验效率，这些均表明本文研究的 psCD-CAT 基本可行，并从方法上真正实现了采用多级评分 CDM 来处理多级评分的 CD-CAT 的思想。

意见 2：文中 2.2 节为“多级评分 CD-CAT 参数估计算法”，实际上讲得是用该模型获取个人参数（个人能力值），而不包括题目参数，这里应表述清楚；

回应：非常感谢专家的建议。的确，在 CAT 或 CD-CAT 中，涉及到的参数估计主要是被试参数的估计，根据专家建议，我们在文章第 12 页修改了相关表述。

意见 3：文中对 CD-CAT 的个人参数估计、选题策略用了很多篇幅来描述，但这方面几乎没有创新的内容，是否可以考虑压缩；

回应：因为本文是第一次提出以 GP-DINA 模型为基础的多级评分 CD-CAT 技术，因此需要对涉及到多级评分选题策略等相关公式进行相关数学推导，以供评审专家审阅其正确性。如果文章有幸被录用的话，我们会根据专家的意见，在终稿考虑压缩相关内容，谢谢专家的意见。

意见 4：研究的实际价值如何体现，作者认为该模型更适合于李克特式量表，可否给出实际的应用例子和数据分析；

回应：本文开发了一种新的多级评分认知诊断模型 GP-DINA 模型，但本文的重点是想探讨该模型下的多级评分的 CD-CAT。当然，本文提出的 GP-DINA 模型是一种新的多级评分 CDM，因考虑到文章的篇幅及本研究的重点及主题，GP-DINA 模型的参数估计精度、性能及其在实际数据应用中的效果我们已另单独成文待发，所以此处未给出 GP-DINA 模型在实现的应用例子，以避免重复，感兴趣读者可以向作者索要。而对多级评分 CD-CAT 的实际运用例子，因为目前我们手中并没有这种类型的大型题库，因此也无法讨论多级评分

CD-CAT 的实际效果，这应该是本文的一个不足，根据专家的建设，我们在文章讨论与展望部分补充了这一不足，详见文章第 22 页，再次感谢专家的建议。

意见 5：论文结构与常见模式不同，第 3 节描述了模拟研究，却在第 4 节和第 5 节分别给出结果；
回应：根据专家的建议，我们将第 4 节和第 5 节合并为一节，即把两个实验的结果都放在第 4 节，并修改了相关描述，详见文章第 18-19 页。

意见 6：论文对结果的描述，以及后面的讨论都显得比较仓促和简短；
回应：专家的建议已采纳，我们对讨论部分做了进一步的补充，详见文章第 21-22 页。

意见 7：文中第 2.1 节中反复出现对涂冬波等人（2010）的引用，其实可以简短截说；
回应：根据专家的建议，我们对相关描述已做了删除及缩减，详见文章第 10 页。

意见 8：写作上尚存在不少小问题，如公式（6）中的“ortherwize”应为“otherwise”，6.1 节的文字以“：”结尾等。
回应：这是我行文时的疏忽，根据专家的建议，我们已修改了公式（6）中有误的地方，并同时修改了公式（8）中的相同错误，详见文章第 11、13、21 页，非常感谢专家的指正。

第二轮

审稿人 2 意见：作者已按审稿意见修改了稿件，同意发表。
回应：谢谢专家的肯定及第一轮修改中提出的宝贵意见。

审稿人 3 意见：作者吸收几位审稿专家意见后，对论文进行了调整修改，进一步明晰了论文所讨论的主要问题，提高了论文的质量和价值。但还需要对以下内容进一步修改：

意见 1：中文摘要需要修改，“本研究对于进一步拓展 CD-CAT 在实践中的应用提供了新方法和技术支持”提供的信息量有限，“拓展出了适合多级评分 CD-CAT（psCD-CAT）的算法与技术”一句也没能充分说清楚作者到底做了哪些工作。
回应：已按专家意见修改，详见文章第 9 页。

意见 2：公式 11 中的 k 应为大写，其他公式中可能也存在类似问题，请作者再次检查。
回应：的确如专家所言，已按专家意见修改，详见文章第 13 页

意见 3：第 14 页公式 15 上方的“计算见公式参见 2.8”存在笔误。
回应：已修改为“计算公式参见 2.8” ，详见文章第 14 页。

意见 4：存在一些细节问题，如“PS-KL 平均使用(10.2+11.14+11.79)/3=11.04）题，”一句中多出的右括号、参考文献列表中的格式和字体。
回应：谢谢专家的细致审稿，根据专家意见我们修改了一些笔误，详见文章第 20 页。

意见 5：建议作者请同行再次阅读，修改其中一些口语化的文字，对表述再次斟酌处理。如第 19 页上“PS-PWKL 和 PS-HKL 选题策略在题库安全性上还有进一步提高的空间，未来研究可以进一步考虑兼顾题库安全性的新选题策略。”一句不仅对表中数据进行分析说明，还增加了展望的意思。而展望部分应放在第 5 节中。而第 5 节的讨论部分请尽量避免再对文章价值进行总结、或重复结论中的话语，以提高论文的整体质量。
回应：根据专家的建议我们再次通读了全文，修改了一些语言表述及表达上的问题，详见文章第 21、22 页，感谢专家的意见。

编委复审：这篇论文对涂冬波（2010）提出的 P-DINA 模型进行了拓广，针对多级评分项目提出了 GP-DINA 模型，这个模型比 P-DINA 模型更合理些。作者还基于 0-1 项目的选题策略 KL、PWKL 和 HKL 给出多级评分的相应选题策略 PS-KL、PS-PWKL 和 PS-HKL。最后通过模拟研究，比较了几种选题策略的效果。该方法对多级评分 CD-CAT 的研究具有推动作用。经三位评审专家审阅，作者对评审专家所提出的问题进行了较好的回答和修订。鉴于三位专家一致同意发表，考虑到这方面的研究对信息环境下个体自适应学习的潜在价值，我推荐发表该论文。

主编终审：可发。