《心理学报》审稿意见与作者回应

题目：基于属性多级化的认知诊断计算机化自适应测验设计与实现

作者：涂冬波，蔡艳

第一轮

审稿人 1 意见：作为一篇具有开创性的文章，本文较好的探讨了属性多级化情况下的CD-CAT 测验特征。比较了较为常用的若干选题策略和终止规则的效果，为后续研究开拓了方向并奠定了一个较好的基础。但仍存在以下问题，希望作者思考。

意见 1.存在笔误。如第一段末尾，qik=0 应说明项目 j 未考察属性 k；2 和 2.1 小标题应为“属性多级化”；公式（2.1）下两行应为“若”等。

回应：谢谢专家的细致审稿及指正，根据专家建议我们修改了些错误，详见文章第 9 页和第 10 页。

意见 2..PA-KL 的效果较差，作者可否分析其较差的原因是什么？是由于 KL 选题法先天上与 PWKL 和 HKL 相比的弱点（即在普通 CD-CAT 中也有类似表现），还是由于 pCD-CAT 多级化的特点造成的？

回应：这主是由于 KL 选题策略本身缺陷所致。KL 选题策略是一种综合性的信息量，即对当前被试估计的知识状态与所有可能的知识状态之间的一个整体的平均区分能力，它并没有充分考虑被试为当前知识状态估计的可能性有多大（如属于当前知识状态的后验概率），也没有考虑当前知识状态与其它知识状态的差异度/相似度，因此对知识状态本身数据的信息利用不如 PWKL 和 HKL 充分。我们又进一步查阅了国内外公开发表的，传统的 CD-CAT 下关于 KL 选题策略的判准率的文献（如 Wang,2013; Mao et al, 2013），发现其判准率均不如 PWKL 和 HKL 高。当然在 pCD-CAT 环境下，由于知识状态更为复杂（取值不仅为 0-1），对知识状态本身的信息利用就显得更为重要，这也进一步导致了 KL 选题策略与 PWKL 和 HKL 的差异加大。因此，我们认为 KL 选题策略一方面有其先天不足，另一方面 pCD-CAT 下属性多级化更为复杂，两者共同导致其在 pCD-CAT 环境下判准率不太理想。

意见 3..本研究对 RUM 模型进行了多级化改造。那么作者为何要选择 RUM 呢？是由于 RUM 使用的广泛性，还是由于 RUM 适合多级化改造？

回应：其实我们对 RUM 和 DINA 两个模型均进行了多级化拓展，拓展的效果均比较理想，但限于文章篇幅及行文表达方便，我们只在该文中报告了 RUM 拓展的结果；当然，RUM 与 DINA 两个模型均是目前认知诊断研究中使用较广泛的模型，相比较而言 DINA 模型更为简洁，而 RUM 虽然相对复杂一点但对被试答对项目概率的数学建模上描述更为精细，有利于进一步揭示及分析被试认知加工过程（de la Torre，2004；Maris，1999）。如果审稿专家对 DINA 多级化拓展的结果感兴趣可以向作者索要。

意见 4.本研究对于 pCD-CAT 的判准率和题库安全性进行了探讨。但作者只比较了不同选题策略对于 pCD-CAT 的影响，而未与普通的 CD-CAT 或习惯性的结果进行比较，因此不能很好的得出 pCD-CAT 判准率较高判准率以及较好的题库安全性的结论。

回应：专家的这条意见非常重要，当初我们进行 Monte Carlo 实验时也想在完全同等条件下比较 pCD-CAT 与传统 CD-CAT，但由于两者的比较不可能完全一样（两者的前提条件不同），前者属性是多级，而后者的属性是 0-1 的。所以我们就改用了文章实验 3 来做一些间接说明，因为实验 3 是探讨当属性为多级化时，用传统 CD-CAT 来处理的危害（这时传统 CD-CAT 的模式判准率不到 30%），当然这也只能从一个倾面来反应开发 pCD-CAT 的必要性及优势。我们之所以得出 pCD-CAT 判准率较高以及较好的题库安全性的结论，不是基于比较得出，而是基于 pCD-CAT 的属性模式判准率平均也高达 0.88（见实验 1 与实验 2，选题策略为 PWKL 及 HKL），以及曝光率、测验效率及测验使用率的指标均较理想的基础上下的结论，文章虽然没有通过相对比较，但通过绝对指标的大小来下结论的。当然，我们觉得专家的意见非常有道理，未来应进一步进行比较研究，为此我们将专家这条意见放在文章最后讨论部分，详见文章第 24 页。

审稿人 2 意见：《基于知识状态-属性多级化的认知诊断计算机化自适应测验设计与实现》一文在传统认知诊断模型的基础上，开发出面向多级属性的 CD-CAT 测验。研究具有创新性，研究方法基本得当。但还有些地方需要修改或再考虑。

意见 1.文章的主要贡献在于拓展原来适合二级属性的模型到多级情景，因此，与原来模型的比较是必须而且是最为重要的，只有这样才能体现出新开发多级属性模型的价值。但目前的

文章中，在仿真实验 3 中才做了此项比较，而且只比较了判准率一个方面的指标、题库安全性、测量精度是否也应该比较呢？同时，实验 3 中比较的是定长（L=20）的情况下两种模型的差异，那不定长的情况是否应该比较新模型与原来模型的效果？建议加强和突出这方面的工作。

回应：专家的这条意见非常有道理。我们当初行文思路是，首先得证明本文新开发设计的 pCD-CAT 是合理可行性，因此才有了实验 1 和实验 2，这两个实验均是在不同条件下来探讨 pCD-CAT 的合理性；只有新开发设计的 pCD-CAT 是合理可行性，那么才可在此基础上进一步开展与传统 CD-CAT 的比较（即实验 3），这是本文的行文思路，我们觉得这个行文思路逻辑是清楚的。但我们觉得专家提出的行文思路也很有道理。我们几位作者经过仔细考虑，如果要将实验 3 与实验 1、2 换个顺序来行文，这样文章结构就进行大修改，因此在不损失表达的逻辑性的基础上，我们没有对文章的行文顺序进行大调整。

综合考虑专家的第 1 条意见，我们对原来的实验进行了修改、补充与拓展：（1）在实验 3 中补充了题库安全性比较（CD-CAT 环境下测量精度主要是指属性的判准率）内容，详见文章第 20 页；（2）在实验 1 与实验 3 中，补充了测验长度为 m=15 题和 m=25 题的两个子实验，即定长 CAT 中，考虑三种定长的规则；（3）在实验 2 中，补充了测量精度（即后验概率）为 p=0.75 和 p=0.85 两个子实验，即不定长 CAT 中，考虑三种测量精度。（4）根据国内外传统 CD-CAT 同类研究，我们调整了题库的容量和被试量，尤其是综合了 Feng 等人 (2014)以及 Hsu 等人（2013）的模拟条件，本研究项目参数调整为$r_{jk}^{*} \sim U\left(0.05,0.4\right)$，$\pi_{j}^{*} \sim U\left(0.75,0.95\right)$，使得项目参数的取值更宽更合理，研究结果的外推能力（或概化能力）提高了。但这些调整并未从根本上改变实验结果及结论。这些修改详见文章第 15-21 页。

实验 3 中由于使用传统方法来处理多级化测量情景，属性的诊断正确率非常低（见文章表 8）；因此在不定长 CAT 中，如果传统方法想达到 p=0.75 甚至 0.85 以上的测量精度，其使用的题量平均超过 100 题（有的被试甚至做完题库所有试题后其测量精度仍未达到要求），这已经失去了 CAT 的意义及价值，因此实验 3 中并没有报告不定长 CAT 下的结果。根据专家建议，我们在文章第 22 页对此做了进一步说明。

意见 2.实验 1 和实验 2 中，作者实际判断的是两种条件下，不同选题策略的差异，这部分内容作为文章次重点更为合适。在实验 1 中，作者比较了判准率和安全性，但定长测验情况下，不同选题策略下的测量精度（后验概率）是相同的吗？如果不同，是否也可以比较？

回应：对于文章行文思路我们仍然认同专家的观点，详见上面回答。实验 1 中，由于实验条件相同，且不同选题策略均使用同等数量的试题，因此如果哪种选题策略在使用同等数量试题条件下其各项指标优于其它选题策略，则说明前者优于后者，这是 CD-CAT 中常用的比较手段。在这里，正因为在相同题量下不同选题策略的测量精度（本研究中测量精度的最后判断标准是被试参数估计的准确度即属性判准率）不同，才会有选题策略的优势之分；同时，实际中也无法指望/保证不同选题策略在相同题量下有相同的测量精度（即属性判准率）。

意见 3.作者三个仿真实验中，只构造了一个题库和被试样本，这样对于题库大小、多级属性数量的多少、样本量的影响没有考察。定长实验中，也只考察了 20 题一种长度。是否可以考虑增加实验条件？

回应：我们认为专家的这条意见非常好。在 CD-CAT 环境下，由于要估计的参数只有被试参数，而影响被试参数估计精度重要影响因素涉及：一是测验长度（定长 CAT）或测量精度（不定长 CAT），另一是使用的选题策略；CAT 中，被试人数的多少会影响项目参数估计精度，但不会影响被试参数估计精度。因此本研究中重点探讨了测验长度（或测量精度）和选题策略这两个因素在 pCD-CAT 环境下的不同表现；同时，限于篇幅、为简化实验设计以及为了更好地说明问题，本研究并没有涉及被试样本量、题库容量等因素，而是采用了二因素实验设计（见实验 1 与实验 2）；对于其它因素（如被试样本量、题库容量等）的影响未来研究需要进一步深入的地方，这应该是本研究的不足之处，为此我们在文章最后讨论部分进行了补充说明（详见文章第 24 页）；根据专家建议，我们补充了测验长度为 15 题和 25 题的测验情景（定长 CAT），详见文章第 18 页；补充了测量精度 p=0.75 和 p=0.85 的测量情景（不定长 CAT），详见文章第 20 页。

意见 4.作者没有交代编程的工具，对新模型所需的计算消耗（计算时间）也没有报告。

回应：谢谢专家的建议。本研究采用 Matlab7.0 语言编程；同时，根据专家建议，我们进一步调查了新模型计算时间：在普通的笔记本电脑下（i5-2450M, CPU 2.5GHz, RAM 2.00G），平均每个被试完成 20 题的 pCD-CAT 用时不到 1 秒，这符合 CAT 的速度要求，当然随着计算机电脑性能的提高以及使用更为优化的语言编程（如 FORTRAN 语言的速度比 Matlab 约快 4 倍），其运算速度还有望提高。

意见 5.参数 r 的取值也来自于均匀分布，而在现实中，该参数是否更符合正态分布？

回应：关于参数 r 及 Pai 的分布，就国前国外研究来看，大多数研究者都采用了先验信息较少的均匀分布，如 Feng 等人 (2014)以及 Hsu 等人（2013）模拟的 r 以及 Pai 均为均匀分布。本研究正是参考了国外这些研究成果。

意见 6.表 3 至表 7 中，列标目都存在很多英文单词，是否应该为中文或英文缩写。其中 3.3.2 节第一句话交代用 ER 作为题库安全性指标，但在表 4 和表 6 中，都改用了过度曝光题目数，前后描述不一致。

回应：谢谢专家的指正，所有指标我们均用中文表达，并删除了一些前后描述不一致的指标。详见文章第 18、20、21 页。

意见 7.表 3、表 5、表 7 中的 A1 至 A5 分别指什么？五道题目的哪个值？表 2 中 a1 至 a5 是用小写字母表示的。

回应：根据专家建议，我们规范了文章前后表达的统一，我们均用表 2 中的 α_1 至 α_5 来表示测验的 5 个属性，即考虑测验 5 个属性每个属性的诊断正确率。详见表 3（第 18 页）、表 5（第 19 页）、表 7（第 21 页）的表头。

意见 8.实验 1 和实验 2 的分析中，都是通过判准率就认为 KL 选题策略不适合，而没有对 KL 策略与其他策略在安全性方面进行比较，既然已经有了相关数据，还是建议综合比较后再下结论更 reasonable。

回应：根据专家建议，我们在实验 1 和实验 2 补充了 KL 选题策略的题库安全性以及测验效率等方面的比较，详见文章第 19、20 页。

意见 9.文章 7 和 8 两节，在部分内容上重复。是否可以考虑合并为 1 节来撰写。

回应：根据专家建议，我们将 7 和 8 节内容进行了合并，详见文章第 22-24 页。

意见 10.英文标题中使用了 with polytomous knowledge states and attributes，这应该是从中文标题翻译而来，考虑国外这方面的文章都只用了“polytomous attribute”或“ordered category attribute”，可否也简化一下？

回应：谢谢专家的建议，我们在文章统一使用“polytomous attributes”的表达，详见文章第 26 页。

意见 11.英文摘要字数没有达到《心理学报》的要求，请补充，并请 native English speaker 检查一下。

回应：根据专家建议，我们补充了英文摘并做了进一步润色，详见文章第 26-27 页。如果文章有幸被录用，我们还会请 native English speaker 帮润色摘要。

意见 12.文中存在一些丢字、错字的现象，请通读全文并修正。例如：P5，"之间，果 πj 越大说明题目越容易，"；P5,"转换公式见 2.2 致 2.4，"；P12，"重叠率越高说明越题库不安全。"；P13，"表 3 说明当采 PA-PWKL 和 PA-HKL 选题策略时，"

回应：谢谢专家的细致审稿，根据专家建议，我们对这些错误进行了修改。详见文章第 10、17、19 页。

审稿人 3 意见:《基于知识状态——属性多极化的任职诊断计算机化自适应测验设计与实现》评审意见文章将属性二级化思想推广到属性多极化，并将 RUM 模型推广到属性多极化的情形，这是本文最典型的创新之处，文中方法可行。另外，在属性多极化的情形下比较了 KL、PWKL 和 HKL 方法的选题表现。但是，文中存在一些错误，例如

意见 1.引言第一段话倒数第四、五个字“掌握”应该为“测量”；

回应：这的确是我们行文时的大意，根据专家建议我们修改过来了，详见文章第 8 页。

意见 2.引言第二段话第五行 α_{jk} 应该为 α_{ik}

回应：已修改，详见文章第 9 页。

意见 3.引言第二段话最后，参考文献的写法有误；

回应：已修改该文献写法，详见文章第 9 页。

意见 4.2.1 部分第二段第一行最后两个字的顺序写反了；

回应：已修改，详见文章第 10 页。

意见 5. .2.2 中介绍的能力估计方法与属性二级化下没有区别啊，实质上文中 2.2 节第三段第一行 $\alpha_l(l=1,2,...,2^K)$ 是不正确，因为属性多极化下，知识状态总数大于 2^K 个，此类错误一直持续到后面，特别是 2.3 节中介绍的选题策略，与属性二级化下除了模型不一样外，没有任何差异；

回应：这是我们行文时的疏忽，的确，在属性多级下，如果不考虑属性间的层级关系，其所有可能的知识状态为 $\prod_{k=1}^{K} L_k$ 种（L_k 指属性 k 的水平数），为此我们在文章中修正过来，详见第 12、13、14、16 页，非常谢谢专家的指正。

意见 6.公式 2.12、2.13 和 2.15 没有必要写的这么复杂吧，你只需要写清楚其中 $P(x_j = x \mid \hat{\alpha})$ 表示项目反应函数就行了呀。

回应：专家的意见有道理，我们简化了公式 2.12、2.13、2.14 和 2.15，详见文章第 13、14 页。

意见 7.3.2.2 中第一段地二行， $P(\alpha_l) = 1/2^K$ 也不适合属性多级化情形吧。

回应：是的，与第（5）个问题相同，我们修改过来了，详见文章第 16 页。

意见 8.文 3.3.1 中公式 3.2 表示的是 MMR，而非 AAMR 吧，并且结果中提到的各个属性的正确判准率的计算公式没有写在这部分中。

回应：根据专家建议我们补充了单个属性的判准率即 AMR，其计算公式详见文章第 16 页。

第二轮

审稿人 1 意见：作者对文章进行了较大幅度的修改，增加了新的实验条件，进一步体现了文章讨论方法的优势，同意推荐发表。

意见:还有两点意见供作者参考：一是将审稿意见回复中交代的编程工具、计算消耗报告在论文中，这是国外期刊同类论文的标准做法。二是对英文摘要做修改，不仅是语言上的调整，也要对内容进行丰富。《心理学报》英文摘要的目标是让国际同行了解我们所做的工作和贡献，建议重点放在 pCD-CAT 的设计思路及其实现上，这也是对国际同行最有参考价值的地方；然后要分别说清楚三个实验的设计和仿真结果，结果可以尽可能用总结性数据说话，同时表明自己使用的具体指标；目前的摘要（尤其是第二段）更像是对文章结构的介绍，实质性的内容还不够。

回应：感谢审稿人提出的两点意见，我们均根据专家意见进行了修改。其中，在文章第 23 页增加了本研究中编程工具、计算耗时等内容；在文章第 27、28 页，进一步补充完善了英文摘要内容。修改了的部分我们均用粉红色进行了标注。

编委复审意见:该研究的选题和设计较有新意，采用的方法科学合理，经过修改后，论文更加完善，所得结论可信，有较好的价值。

主编终审意见: 一些格式的小问题希望编辑部把关，如：Wang，Chang & Huebner；Hsu, Wang & Chen, 2013；Feng，Habing,和 Huebner（2014）。