《心理学报》审稿意见与作者回应

题目：基于属性掌握概率的认知诊断计算机化自适应测验选题策略
作者：罗照盛 喻晓锋 高椿雷 王睿 李喻骏 彭亚风 王钰彤

第一轮

审稿人 1 意见：《基于属性掌握概率的认知诊断计算机化自适应测验选题策略》提出在 CD-CAT 选题策略中使用属性掌握概率代替属性掌握模式估计值，并且提出了六种可行的选题方法，有一定创新。以下问题请作者参考。

意见 1：在本文的引用文献中，辛涛，刘拓（2013）总结了以提高测量精确性、控制题目曝光率、平衡测验内容为目的的三大类 CD-CAT 选题策略。在本研究的模拟研究中作为参照的传统选题策略 KL，SHE，PWKL 和 HKL 都属于以提高测量精确性为目标的选题策略。研究者所提出的六种选题策略是否也是以提高测量精确性为目标呢？希望作者能够在文章明确表示。

回应：在本研究中，提出了几种基于属性掌握概率的选题策略，作为新提出的选题策略，首要的是研究它们的测量精确性，在此基础上，进一步研究其在题目曝光率和题库使用均匀性等方面。正是基于此，本文对基于属性掌握概率的选题策略在测量精确性、题目曝光率、题库使用均匀性等方面都进行了研究，并且与常用的几种策略进行了比较。本研究提出的选题策略希望能综合考虑精确性，题库使用均匀性等方面的因素，使选题策略具有实用性。我们补充了对研究新策略的目的的相关说明。

意见 2：作者采用了几个题库使用均匀性的指标来比较不同的选题策略。如果选题策略主要是为了提高知识状态估计准确性，那么题库使用均匀性仅仅是一个副产品。请作者对题库使用均匀性的比较进行解释和说明。

回应：新选题策略的测量精确性是一个非常重要的方面。在实际的测验开发中，题库的使用均匀性也很重要，我们对此增加了解释和说明。

意见 3：虽然二者的计算方法不一样，后验概率与属性掌握概率有很多信息是重叠的。希望作者对这方面有所解释。

回应：感谢专家细致的审稿。基于属性掌握概率的选题策略的基本过程是：(i)先随机生成被试的属性掌握模式；（ii）然后用选定的选题策略选取项目供被试作答；（iii）基于被试的属性掌握模式(随机生成或估计值)、作答计算被试属于每种属性掌握模式的后验概率、被试在每个属性上的掌握概率（记为 P1）；(iv）基于当前的属性掌握模式估计值和已作答过的项目，从剩余题库中选择一个项目供被试作答，计算被试对每个属性的掌握概率（记为 P2）；（v）按照预定的准则决定选择下一个项目，比如 P1 和 P2 之间的差值最大（即最大属性掌握概率变化法）等。后验概率体现的是基于已经作答项目的条件下被试属于每种属性掌握模式的可能性，而属性掌握概率一方面体现了被试基于已经作答项目的条件下被试对每个属性掌握的可能性(P1)、另一方面它也体现了被试基于已作答项目和下一个可能项目的条件下对每个属性掌握的可能性(P2)。因此，虽然后验概率与属性掌握概率之间有很多信息是重叠的，但是考虑了属性掌握概率的选题策略会更精细的体现出不同项目对于被试的影响。

意见 4：在摘要中，作者提出“模拟实验结果表明，基于属性掌握概率的选题策略有很好的

分类准确率，在题目曝光率，题库使用均匀性等方面也有很好的表现”，似乎有些夸大。因为根据表 1 和表 2 的结果，本研究提出的六种新方法似乎只有 PPWKL 有比较满意的表现。作者在文中并没有对此进行讨论。
回应：我们的表述不够严密和准确，对这部分的内容进行修改，以使表达更准确。

意见 5：从表 1 的结果看即使 PPWKL 的表现看似略优于 PWKL 和 HKL，但 PPWKL 的优势并不明显：PPWKL 的模式判准率略低于 PWKL 和 HKL，最大曝光率这个副产品仅仅从 0.97 左右降到了 0.938，未使用项目数减少了不到 2%，这些微小的差异是否有重要价值值得探讨。表 2 也有类似的情况，PPWKL 似乎有一些微弱的优势。
回应：如果单从判准率上看，PPWKL 的优势似乎并不明显。本研究中提出的选题策略希望基于实际应用的视角，即要有较高的判准率、在题库的使用均匀性等也综合考虑。我们增加了相关的讨论。

意见 6：模拟研究二进一步为 PPWKL 的“优势”提供了证据，但似乎有些牵强。首先，当测验长度小于 6 时，各个选题策略的模式判准率都低于 0.5，在这种情况下比较谁的模式判准率孰高孰低似乎没有太大的实际意义。本文也引用了 Chun Wang 2013 年的文章，建议作者参考一下 Wang(2013) 文中短测验的模式判准率。
回应：我们接受专家的建议，在修改稿中增加了与 Wang(2013)文中所介绍到的互信息选题策略的比较。并对结果进行了讨论。

意见 7：作者在文中反复强调“在测验初期，…，采用基于属性掌握模式的选题策略可能不利于被试的知识状态的估计”。然而，虽然测验初期属性掌握模式估计不准，但至少从表 1 的结果来看，基于属性掌握模式的几种方法（除了 KL）在测验结束时都提供了比较高的属性判准率。希望作者能对此进行解释和讨论。
回应：已有的几种典型的 CD-CAT 选题策略在对被试的分类精度上已经做的很好，这在很多文献中都有提到。本研究中提出的选题策略在精度上略好，并且在测验较短时优势更大，这就对基于本文中的策略进一步考虑非统计约束提供了更大的操作空间，我们增加了相关内容的解释和讨论。

审稿人 2 意见：认知诊断计算机自适应测验是一个近年来新兴的研究问题。《基于属性掌握概率的认知诊断计算机化自适应测验选题策略》一文能够考虑到基于被试属性掌握模式的估计值进行选题可能造成的偏差，从属性掌握概率来进行选题，这是一个新的视角，从新的方面来解决问题，具有一定的创新性，很值得学习。通读全文，还有以下不妥之处，请作者予以修缮。
意见 1：摘要部分阐述不够清晰，研究结论表述不妥，请对摘要进行精细修练。
回应：根据专家的建议，对摘要部分和研究结论进行修改。

意见 2：文中从 CD,CAT 一一道来，再说明 CD－CAT。但关于 CD－CAT 的优势作者阐述的还不是很足，请结合已有的相关实例或参考文献进一步说明 CD－CAT 的优势。
回应：我们补充了对 CD-CAT 的介绍和说明，进一步介绍 CD-CAT 的特点和优势。

意见 3：请作者针对这几个概念再进行详细说明:1）被试属性掌握模式；2）被试属性掌握概率；3）被试属性掌握概率变化加权；4）属性掌握模式距离加权等相关概念。

回应：补充对专家提到的概念的介绍和说明。

意见 4：“基于属性掌握概率的选题策略是从属性掌握概率出发，对属性掌握概率不作 0、1 转换，选择对被试属性掌握概率影响最大的项目作为下一个施测的项目”，为什么是影响最大的项目,请做出解释与说明.
回应：诊断测验的目的是根据被试在项目上的作答推断出其属性掌握模式。属性掌握模式是由 0，1 组成的向量，并且它是根据诊断模型计算出的属性掌握概率转换而来的，如果采用 0.5 作为截断点，比于属性掌握概率 0.1 和 0.4 都会转化成 0，但是从统计上来说，0.1 对应的属性掌握情况更可能是未掌握（即 0），因此，在选择项目时选择对属性掌握概率影响更大的项目作为下一个施测的项目更有利于估计被试的属性掌握模式。我们增加了想说的解释和说明。

意见 5：实际应用当中题库结构相当复杂，但本模拟研究仅讨论了属性间相互独立的关系。请作者对其它的层级关系进行讨论，如果不展开相关讨论，请予以说明为什么不展开。
回应：我们增加了对属性层级关系的相关讨论。

意见 6：文中提到“与陈平等(2011)相同”，为何要强调这一点，是为了比较，还是为了研究的简便。如果为了比较，请在研究结果与结论当中进行体现。
回应：我们采用模拟数据的方法与陈平等（2011）相同，主要是为了研究的简便。

意见 7：“3.3 评价指标”提供了许多的评价指标，但都是逐个进行比较，由于涉及的指标较多，最后难以权衡。请提供一个或两个综合评价指标从准确性和题库使用均匀性两方面来展开综合比较。
回应：我们提供了相应的综合评价指标来综合比较相应的选题策略。

意见 8：“3.4 实验结果”由于涉及的选题策略和评价指价众多，阐述的层次性不够清晰。请从已有的 CD－CAT 选题策略和基于 AMP 的选题策略两大方面进行比较；然后再逐项比较等。
回应：根据专家的建议，对 3.4 实验结果进行重新改写，力图使介绍更清晰。

意见 9：“3.4 实验结果”方面有以下几个疑惑：1）仅从排列名次上分析,第 1 和第 2 的值有时相差很小，这种差异是否存在显著性.比如知识状态准确性方面 HKL 和 PWKL 这两种策略分别是 0.833 和 0.838,相差 0.005,该差异值很小,是否具有显著性呢? 2）从所列的指标来看,有些准确性较好,但另一方面题库使用均匀性方面却不是很理想.反之亦然.例如,AMP1 选题策略,在题库使用性方面有两项排第一,但是在准确性方面却不理想.请给予解释.并说明这种现象是普遍的,还是典型的,并进行深入讨论.3）从现有结果来看,作者所提到的 6 种选题策略经比较,只有一种尚可,而已有的 4 种选题策略当中,却有两种都不错.在定长的条件下,PPWKL 的优势并不突出.请作者予以解释.
回应：感谢专家细致和耐心的审稿，本研究中的实验结果都是在实验多次重复之后，计算的平均值，因此，可以认为，实验结果是排除了随机误差的。从结果来看，排名在前几位的指标之间确实差异较小，单从对被试的估计精度来看，已有的 PWKL，HKL 等策略已经做的很好，这一点从众多的研究 CD-CAT 选题策略的文献中可以看出，如果考虑题库的使用等指标，本文中涉及到的选题策略都有很大的改进空间。
本文中提出的 6 种策略中有一种策略，即 PPWKL 在估计精度上略占优势，并且综合估计精

度和题库使用等指标后，PPWKL 在文中提到的选题策略中排名第一。
专家提到在定长的测验中，PPWKL 优势不突出，这个是因为随着测验长度的增加，被试的属性掌握模式的估计会慢慢趋近于其真值，在这个过程中，PPWKL 策略的优势会变小。
实际上，估计的准确率和题库使用的均匀性是相冲突的一对指标,第一，虽然，在选题策略中的确存在两者都比较好的情况，如 PWKL 在两个指标上都比 KL 好(Xu, Chang, & Douglas, 2003; Cheng, 2009)，在这种情况下，较差的那个选题策略的实用性就很低了；第二，同一个选题策略，如果估计精度很高，曝光控制一般不好，比如 CAT 中最大信息量选题方法；按 a 分层估计精度稍微差一点，但是曝光均匀性大大改善；第三，在实际的应用中，应该针对测验的目的，综合权衡。
针对专家所提到的问题我们在文中都增加了补充说明。

意见 10：第“5 小结与讨论”部分，请对研究结论进行清楚阐述；其次请再结合研究结果进行深入讨论，为何 HKL，PWKL 和 PPWKL 这三种选题策略要优于其它的选题策略；虽然 PPWKL 相比较而言，有优势，但并不突出等，请予以深入讨论。
回应：我们增加了对相关内容的讨论和说明。

审稿人 3 意见：本文的核心内容在于探讨基于属性掌握概率的选题策略的表现，并提出了 6 种新的选题策略。整体审稿意见如下，详情参见审改稿，仅供作者参考：
意见 1：6 种新选题策略名仅给出了缩写，建议给出英文全称；
回应：根据专家的建议，增加了新策略的英文全称。

意见 2：作者的不定长终止规则仅考虑了“属性掌握模式后验概率最大值固定为 0.8(Tatsuoka, 2002)”这一情况，而没有考虑属性掌握模式的后验概率最小值的情况，请参见 XX()一文。这点是否修改，仅供作者参考，因为终止规则的改变必然会影响模拟研究结果。
回应：感谢专家的提醒，终止规则的改变一定会影响模拟研究结果。在不定长终止规则中，属性掌握模式的后验概率表明了测验对被试属性掌握模式的确定性程度，在一些文献中，通常采用 0.8 作为界限，即只要被试对某个模式的后验概率达到 0.8，即判定其属于该种模式。我们也进行了采用其它后验概率(比如 0.6)的方法，从结果来看，对被试的属性分类准确性没有采用 0.8 时好。

意见 3：属性掌握概率和属性掌握模式的差别在于，“概率”将各个属性独立开来(即未考虑属性间的交互作用)，而“模式”是一个整体(考虑到了属性间的交互作用，参见 GDINA 或 LCDM)，所以将两者进行比较的前提是假设各个属性间相互独立。而作者在研究中也是采用“被试对每个属性的掌握概率按 0.5 进行模拟”的，所以本研究的结论或许是有局限性的。建议作者进行额外探讨，尝试探讨下当属性间存在交互作用时采用“概率”好，还是采用“模式”更好。
回应：感谢专家细致的审稿，属性间存在相关和属性间存在层级关系时新选题策略的研究非常值得进一步研究。我们在文中增加了与另外一种选题策略(基于互信息的选题策略， 可参见 Wang，2013)的比较。而各选题策略在属性间存在相关或层级关系时的比较涉及的内容也较多，故我们放在讨论中进行说明。

审稿人 4 意见：在涉及到 KL 信息量的选题策略中，还需要估计被试的属性掌握模式。也就

是说在 PPWKL 等几种选题策略中，既需要使用 MAP 估计被试属性掌握模式，也需要使用 EAP 估计被试属性掌握概率，这两种方法的估计结果不一致时以哪种方法为主？所考察的后三种选题策略与具体哪一种估计方法的不同结合是否会导致不同的结果？

回应：感谢专家细致和耐心的审稿，我们对相关的问题进行了解释和说明。从目前我们所掌握的文献资料来看，对被试属性掌握模式的估计过程大多是采用 MLE、MAP 或 EAP 来估计被试的属性掌握模式。当被试总体的属性掌握模式按均匀分布考虑时，MLE 与 MAP 是等价的(Huebner, Wang, 2011)，MLE、MAP 分别是计算被试属于某种理想模式时的似然函数最大和后验概率最大，进而选择该种理想模式作为被试的属性掌握模式估计值。EAP 计算被试的属性掌握模式要经过三步，首先要采用 MAP 计算出被试属于每种属性掌握模式的后验概率；然后再转换成每个属性的边际概率；第三步是将属性掌握概率用截断点的方式转换成 0，1 组成的属性掌握模式。根据 Huebner 和 Wang 在 2011 年的研究结果，表明这三种方法在属性掌握模式、单个属性等的估计精度上没有大的差异，它们各有各的优势，比如单就属性掌握模式的估计精度来说，MAP 稍微占优；但是考虑单个属性的估计精度来说，EAP 更好，并且 EAP 更少会出现将被试的 K 个属性全部估计错误或 K-1 个都估计错误的情况。

我们在程序中采用了 MAP 估计被试的属性掌握模式，采用了 EAP 估计被试的属性掌握概率，这其实并不矛盾，因为 EAP 估计被试的属性掌握概率时用到了 MAP 所估计的属性掌握模式，采用 EAP 的目的只是为了计算被试的属性掌握概率(被试的属性掌握模式采用 MAP,不采用 EAP,因为从模式估计上看，MAP 略有优势)，因此，不会出现二者结果不一致的情况。

意见 1：PPWKL，PSWKL,PUWKL 都是在 PWKL 方法基础上的加权改进，那么一个预期的结果就是这三种新方法或者其中某一种会优于 PWKL 方法，否者所进行的加权是没有意义的，只能增加计算量。但是实验数据中所看到的结果是，这三种方法中的效果最好的方法 PPWKL，与 PWKL 方法相比，在判准率和题库调用均匀性方面并没有显著提升。虽然单纯从卡方指标来看，PPWKL 稍有提高，但是在未使用项目数，曝光超过 20%的项目数两个关键指标上并没有明显改进，也就是新方法在题库调用均匀性上并没有看出明显优势。

回应：如果看单个指标，基于属性掌握模式的选题策略和考虑了属性掌握概率的选题策略之间是各有优势。为了能综合地比较这两类选题策略之间的优劣，我们采用了比较选题策略时常用的统一量纲的加权求和的方法，结果显示，从综合指标上来看，考虑了属性掌握概率的选题策略 PPWKL 更好。

意见 2：在短测验中，在一定程度上揭示了 PPWKl 方法与 PWKL 方法运行特点的一些差异，但两种方法的比较最关键是要看最后达到稳定估计结果时的差异。虽然 PPWKL 在前面几个题中确实稍高于 PWKl 等方法，但是在前面几个题目中这种判准率整体是很低的。那么在题目较少时，判准率较低情况下，新方法的优势是否有使用价值呢？

回应：感谢专家细致的审稿，在选题过程中考虑被试的当前能力估计值(属性掌握概率)在目前来说，是一种探索。从目前的实验结果来看，将属性掌握概率考虑进来，会对整个测验产生影响，这一点从统一量纲后的综合指标可以看出。当然，这两类选题策略（基于属性掌握模式的和考虑了属性掌握概率的策略）之间似乎差异不大，这里面可能的原因是这两类选题策略在设计上考虑更多的是对被试的分类准确性上，因此，即使考虑进了属性掌握概率，对被试的分类精度提高的幅度并不大。在 CD-CAT 中，考虑题库的使用均匀性和测验的安全性已经被研究者们重视，将题库的使用和测验的安全性与本文中提出的策略相结合应该是有所作为的，我们在讨论中对此进行了阐明。

第二轮

审稿人 1 意见： 感谢作者对评审意见作出的的积极回应。以下小问题请作者参考：
意见 1：建议标题中不要用缩写，除非是非常常用的（如 DNA、IRT）。“2 基于 AMP 的选题策略”中的 AMP 可能是 probability，也可能是 pattern，容易引起歧义。
回应：已经根据专家的意见进行了修改。

意见 2：英文摘要中的一些用词值得商榷。例如“during CD-CAT, these familiar methods would use a cutoff point to transfer...” 中的 during 和 transfer, do you mean "transform"?
回应：感谢专家细致的审稿，这里用 ”transform” 更合理，我们对英文摘要进行了检查和修改。

意见 3：还有第二段第二行的“the first strategies selects...”
回应：我们对全文中涉及到的英文表述进行了检查和修改。

审稿人 2 意见：基于现有 CD-CAT 选题策略，作者从属性掌握概率的角度，提出六种新的选题策略，这对探索 CD-CAT 选题策略问题无不是个很好的新视角。作者对每位专家审稿意见都做了详细、流畅的回复与修改。再次阅读全文，还是有以下几点疑惑。
意见 1：作者对文中所涉及到的若干种选择策略进行了模拟比较研究，但是从研究结果的各项评价指标值来看，这些选题策略间的差距很小，几乎无显著性，而且作者又是在很特殊的属性层次关系（如独立性）下获得的结果，这不免让人觉得这些差距存在的现实性。
回应：感谢专家耐心和细致的审稿。在文章的修改稿中我们增加报告了关于属性间存在相关的情况，改成属性之间存在低、中和高相关时，比较各选题策略的表现。另外，在原文中一共有 6 种基于属性掌握概率的选题策略，但是并不是表现都好，考虑文章篇幅的问题，因此在修改稿中只保留表现较好的两种基于属性掌握概率的策略。

意见 2：其次，通过评价指标值的排序与综合指标的比较，发现 PPWKL，PWKL 和 HKL 三种选题策略是比较好。请作者思考并阐述为何这三种策略是比较好的。
回应：我们根据专家前面的意见修改了实验设计，对相关结果增加了相关的讨论。

意见 3：虽然 PPWKL 是最好的选题，但是从指标值来分析它并没有很大的优势，尤其是它与 HKL 策略相比较。前者是基于后验概率和属性掌握概率加权的 KL，后者是基于后验概率和属性掌握模式距离加权的 KL。由此设想到，PPWKL 或许做为初测的选题策略较好，因为随着被试作答项目的增加，对被试属性掌握模式估计的精度增加，PPWKL 也就体现不出它的优势了。关于这点，请作者加以探讨。
回应：正如专家所提到的考虑属性掌握概率的选题策略在测验初期对被试的知识状态估计精度有改善，在测验后期，对被试的知识状态估计精度与其它策略相近。在整个测验阶段，考虑属性掌握概率的选题策略都使得题库使用均匀性都有改善，这一点类似于传统 CAT 中在不同测验阶段使用全局信息量和 Fisher 信息量之间结果的比较。我们对此进行了讨论。

审稿人 3 意见：审稿人认为这个“最不确定属性”的逻辑是有错误的。根据作者的定义“最不确定属性掌握概率是指与 0.5 最接近的属性掌握概率”，那么有没有一个范围呢？

意见 1：比如某 α=[0.45,0.49,0.48]，则根据作者的定义，会认为属性 2(0.49)是“最不确定属性”，而属性 1(0.45)并不在作者考虑的范围里。那么另一个 α=[0.45,0.69,0.98]，这时属性 1(0.45)又成了“最不确定属性”了。

回应：感谢专家细致的审稿，这里所提到的最不确定属性是指属性掌握概率最接近 0.5 的那个属性。正如专家所举的例子，在 α=[0.45,0.49,0.48]中，第 2 个属性是最不确定属性，我们选的下一题会使其掌握概率变化最大；而对另一个被试 α=[0.45,0.69,0.98]，选择的下一题会使第 1 个属性的掌握概率变化最大。“最不确定属性”只是个相对的概念，只是其对所有属性掌握中相对不确定的那个属性，我们的出发点是“如果项目能最大提高对最不确定的属性的掌握情况的判断，则该项目是一个较好的项目”。

意见 2：另外，当再有当 α=[0.95,0.89,0.98]时，“最不确定属性”的属性掌握概率是 0.89，都接近 90%了，还“不确定”吗？

回应：我们通过实验表明，当 α=[0.95,0.89,0.98]时，这时被试的属性模式已经“很确定”了，达到这个状态，可能已经达到测验的终止规则（最大长度或模式后验概率达到 0.8）了。当然，如果此时测验还未满足终止规则，下一题会选择使第 2 个属性的掌握概率变化最大的题，会进一步提高被试对该属性掌握程度的判断。

根据后面的实验结果也能发现该方法并没有带来什么优势。

意见 3：作者在上文提到“因此考虑研究基于属性掌握概率变化最大的策略的表现，目标就是尽快确定这些“非常不确定”的属性的掌握情况。”似乎行文逻辑存在问题。

回应：我们的表述不够严谨，文中提到的基于属性掌握概率变化最大的选题策略有好几种：分别是基于属性掌握概率之和变化最大，最不确定属性掌握概率变化最大以及单个属性掌握概率变化最大。我们修改了表述，使之尽量能准确表达。

意见 4：作者的不定长终止规则仅考虑了“属性掌握模式后验概率最大值固定为 0.8(Tatsuoka, 2002)”这一情况，建议作者参考关于 CDCAT 不定长研究的最新进展：

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. Applied Psychological Measurement, 37, 563-582.

新的研究应该基于新的研究成果之上.

回应：根据专家的意见，我们参考了 Hsu 等（2013）中的相关研究成果。

意见 5：作者原文中指出“被试属性掌握模式的模拟与陈平等(2011)相同，即题库中的项目数为 360”，题库设置的太大了，体现不出各个方法的差异。根据后面的实验结果也能看出，有约 2/3 的题目都没有适用。这样会弱化各个方法的差异。建议将题库设定为 200。审稿人认为研究不是图方便，“与 XXX 一致”的做法往往需要作者后期检验这种图方便会对研究结果带来什么结果。

回应：根据专家的建议，我们增加了随机生成题库（200 个项目）时的研究，并且将实验设计修改为属性之间存在不同程度的相关（较低，中等和较高）时各选题策略的比较，报告了各策略的表现。

意见 6：“3.2　CDCAT 测验施测过程”中作者写到“(3) 模拟被试作答；(4)采用 EAP 方法估计被试的属性掌握概率。”，建议作者扩展该部分内容，并提醒作者，如果作者从精细角度去思考选题策略那就应考虑到从精细角度去探讨参数估计的精准度。

回应： 根据专家的建议，我们增加了对 EAP 方法的详细介绍。

意见 7：统一量纲存在不公平性。因为测验精度只有 PMR 和 MMR 两个指标，而题库均匀性有 7 个指标，因此，在等权重的同一量纲并不公平。建议精度指标权重设定为 7/2，而均匀性指标权重设定为 1。或其他公平性权重。不等权重后排名不一定会产生变化，但会增加公平性

回应：根据专家的意见，我们采用了不等权重的统一量纲计算方法。

意见 8：研究二逻辑不通。Wang 等仅讨论了当属性间存在较高相关时 MI 比 PWKL、KL 和 SHE 好，那么她不讨论属性间相独立是有原因的：因为在现实情境中，属性间存在相关性的情况一定是多余相互独立的情况的(甚至可以说不可能存在相互独立的属性)。所以正常的研究逻辑是：由于技术和方法的限制，先假设属性间相关为 0 去进行探讨。之后随着研究的进展应讨论更符合现实的属性间存在相关的情况。但作者的实验二逻辑是，在已经有讨论属性间为相关时的情况下，再去讨论属性间相关为 0 的情况，虽然得到了不一样的结果，但与 Wang 在属性间存在相关去探讨的研究相比，该结果和结论是“无意义”的。

而且，实验二的设计并不完善，如果作者需要那 PPWKL 和 MI 进行对比，则应该有较为公平的对比起点，即应该考查至少两种情况：1.属性间相关为 0；2 属性间相关为 x。如果条件允许，还是应该将 x 大致分为低相关、中相关和高相关。

所以，从实验设计的出发点和逻辑上讲，审稿人并不支持该研究的实验二中与 MI 的对比研究。

审稿人认为，导致作者仅探讨属性间相关为 0 是由于本文的核心概念(属性掌握概率)的缺陷导致的。和一审意见一致，把属性单独出来进行参数估计和以属性模型向量进行参数估计的结果可能是不一样的，尤其是当属性间存在相关时。

回应：正如专家所说的，在现实的测验环境中，属性之间很可能是存在相关的。为了使模拟情形更接近于实际情况，我们修改了实验设计，模拟了属性之间存在不同大小的相关时，各选题策略的表现。

意见 9：或许审稿人并没有讲清楚，审稿人所指的以属性模式向量进行参数估计是指作者应考虑属性模式向量的后验概率，而不是单个属性的后验概率。比如 K=3，那么应该有 8 种属性模式，则参数估计时，作者可考虑在 8 个节点上进行选择，最后统计每个节点的后验概率，那么这应该就是属性模式向量的后验概率。与单个属性概率累积所不同的是，属性模式向量在参数估计时候已经将属性之间的相关性纳入其中了。

回应：感谢专家细致的审稿，这可能是之前我们没有表达清楚。本文修改稿的两种基于属性掌握概率的选题策略 PPWKL 和 PHKL，因为都涉及到后验概率加权，因此考虑了可能的属性掌握模式的后验概率，以及对被试在各个属性上的属性掌握概率所带来的变化。其实我们在模拟实验中，考虑了基于被试的属性掌握模式后验概率变化最大的方法，只是从结果来看，这种方法的效果不算好。

审稿人 4 意见： 统一量纲的综合指标具体如何计算的。测验精度和题库调用均匀性是两个不同的问题，这样统一量纲的依据在哪里？效果最好的 PPWKL 选题策略，未看到有明显

优势，判准率方面 PWKL 选题策略的效果本身就比较好，在此基础上加权的 PPWKL 没有明显提高，题库调用均匀性方面新方法并没有从根本上解决这个问题，如果仅仅是数据变换形式后的呈现，那应用的价值在哪里？
新方法提出的出发点到底是要解决什么问题？是想提高判准率，还是改进题库调用均匀性，似乎并不明确。
以上仅仅是个人意见，难免偏颇。在阅读这篇文章的过程中，也有很多让审稿人学习的地方
回应：根据专家的建议，我们在文中增加了统一量纲的详细计算过程。
另外，关于本研究的出发点或研究的价值:（1）希望从一个不同的角度和路径来探讨 CD-CAT 的选题策略问题，也许可以对后续的其他研究起到启发的作用；（2）本研究所提出的选题策略在测验初期有一定的优势，这个发现是否可以启发类似传统 CAT 测验中采用 Global Information 和 Fisher Information 组合选题的思考和进一步研究；（3）从实践角度来说，教室中的随堂测验或学生的日常自我诊断测验可能可以大量使用短测验，当然这个问题还需要更多的综合研究；（4）从模拟结果来看，相对于典型的策略如 PWKL，SHE 等，新的选题策略在不损失或较少损失测验精度的条件下，能在题库使用均匀性上有改善。这也是构建新的选题策略的初衷之一，即一方面有较好的测验精度，同时有较好的题库使用均匀性，即构建一个在综合指标上表现较好的选题策略。这些考虑是否恰当还请审稿专家指正。

第三轮

审稿人 1 意见：作者较好地回答了审稿人的问题，本次修改稿质量高于之前的。
回应：感谢专家的肯定。

编委复审意见：
意见 1：文章请压缩至 1 万字以内（包括表格及文献）。
回应：根据编委的意见，在不影响意思表达的前提下，我们对全文内容进行了调整和删减。

意见 2：现在有 3 个附录，建议：
（1） 保留附录 1；
（2）附录 2 和 3 共有 7 个表格，请将 7 个表格的结果归纳为文字，300 字差不多吧！
回应：根据编委复审的意见，我们将只保留了附录一，删除了其它三个附录，并在正文中标明：如果需要，可以与作者联系。对删除的三个附录中的结果用文字进行了描述。

意见 3：针对专家对文章摘要的的意见。
回应：谢谢专家的指导，我们找了一位在美国的学者对英文摘要进行了修订，当然，如果专家能够进一步给我们指导，我们当感激不尽！