《心理学报》审稿意见与作者回应

题目：基于基尼指数的双目标 CD-CAT 选题策略
作者：罗芬，王晓庆，蔡艳，涂冬波

第一轮

审稿人 1 意见：

意见 1：本文提出了进行双目标认知诊断计算机化考试的新算法——基尼算法，通过模拟研究展示了新算法的一些优越性，具有一定的现实意义。现提出以下问题供作者思考：

介于 DWI 方法在定长 20 题中的 PMR 指标不佳，建议更换为 zheng 等人（2018）提出的最新变式 PWCDI*KL，或者 PWACDI*KL。zheng and Chang (2016)中显示 PWCDI， PWACDI 比 SHE 法更加高效，因此有可能使用 PWACDI*KL 会改进表现。

回应：我们按照您提供的意见，在第二稿中的模拟实验中用 IPA 策略中的代表 PWACDI*KL 策略代替了 DWI 策略。先在预实验中，将 PWACDI*KL 策略与 DWI 策略进行比较，以 G-DINA 模型为例，模拟 1000 个被试，具体结果如下表，实验结果表明， PWACDI*KL 策略的测量精度确实优于 DWI 策略。因此，本研究最后补充报告了 PWACDI*KL 策略的实验结果。

评价指标	PWACDI*KL			DWI		
	HO	属性间高相关	属性间低相关	HO	属性间高相关	属性间低相关
PMR	**97.3%**	**97.8%**	**97.1%**	73.6%	94.9%	90.9%
Bias	0.07	-0.02	0.08	0.01	0.01	0.04
RMSD	0.28	0.35	0.28	0.47	0.31	0.33
χ^2	75.77	63.05	81.04	89.78	65.38	89.14
TOE	0.37	0.33	0.40	0.43	0.34	0.44
TC	23.40	24.07	23.49	2.11	2.09	2.12

意见 2：不知 theta 的节点数是否影响本文的研究结果？

回应：我们认为 theta 的节点数对研究结果有一定的影响但不会太大，因为算法部分关于 theta 信息量的计算中，theta 的置信区间是确定的，只要节点数不是太少，积分运算的拟合就可以满足要求，当然节点数多，拟合程度会更好，但同时会因运算量增大而增加选题时间。根据 BILOG 程序中计算后验期望概率的推荐，本文中的求积结点数取与$2\sqrt{t}$（t 为已做答项目数）相近的自然数。测验初始时，参数估计不够准确，节点数可以少些，随着测验长度的增加，节点数也会相应增加，最大节点数为 9。为此我们做了如下比较实验，将节点数固定为 32 个，实验结果表明节点数的增加能少许提高测验精度。实验条件为 G-DINA 模型下属性间高相关，模拟了 1000 个被试，结果如下：

	$2\sqrt{t}$个节点						32 个节点					
	PMR	Bias	RMSD	χ^2	TOE	TC	PMR	Bias	RMSD	χ^2	TOE	TC
Gini	97.22%	0.00	0.29	69.37	0.36	2.27	98.3%	0.00	0.35	55.21	0.30	2.63
ASI	93.05%	0.01	0.29	77.30	0.39	0.82	93.1%	0.02	0.32	66.02	0.34	1.20
IPA	97.44%	0.02	0.29	77.11	0.39	21.95	97.4%	0.02	0.35	57.40	0.30	23.82
JSD	92.02%	0.02	0.30	53.26	0.29	0.16	91.5%	0.05	0.34	61.81	0.32	0.15

意见 3：在没有使用项目曝光的前提下讨论题库均匀性是否具有实际意义？第一，模拟研究中呈现的卡方值是否具有实际的差异？第二，使用项目曝光之后，这些差异是否会消失（当然存在一个是否需要使用曝光技术的问题）？

回应：我们认为没有使用项目曝光的前提下讨论题库均匀性有一定的实际意义。在国外 CAT 研究中(Chang, Chiu, & Tsai, 2019; Cheng, Diao, Behrens, 2016; Dai, Zhang, & Li, 2016; Kang, Zhang, Chang, 2017; Liu, Cai, Tu, 2018)中，也常会讨论在没有使用项目曝光控制的前提下讨论题库均匀性，因为新策略与传统策略相比，如果测量精度相当，但题库利用均匀性更好，那这种新方法更具优势，也更值得推荐使用。

模拟研究中如果各方法间卡方值差异很小，在实际测验中可能没有差异，如果差异足够大，则会有实际的差异。卡方值和 TOE 都可以用来衡量题库利用的均匀性，且两个指标的变化是一致的，TOE 值代表两个被试作答相同项目的比例，我们以 TOE 值为例：在表 3 中，G-DINA 模型下属性间低相关，Gini 策略的 TOE 值为 0.37，ASI 策略的 TOE 值为 0.44，两者差异 0.07，这意味着用 Gini 策略选题相比于用 ASI 选题，选择到的相同项目平均会减少 0.07*20=1.4 题，即能减少 7%的相同题量，这有利于增强测验的安全性。

对于实际高风险测验，曝光技术的控制是十分重要的。测量精度和题库利用均匀性是一对相互冲突的指标。使用控制项目曝光技术后，题库利用均匀性会更好，但也会带来测量精度下降的不利影响，如何权衡需要在实际应用场景中根据需要进行选取。另外，使用控制项目曝光技术后，方法间的这些差异是否会消失，有待进一步研究，谢谢专家给我们提供了新的研究设想，我们对这一问题在文章的讨论部分进行了补充说明。

审稿人 2 意见：

意见 1：“一种基于不确定性度量指标的双目标 CD-CAT 选题策略”一文围绕自适应认知诊断选题策略的研究，提出了一种基于基尼系数的双目标选题策略，并采用两个模拟研究探讨了选题策略的实际效果。文章总体上具有一定的创新性。但是就文章目前的写作和结构，尚有以下需要修改和进一步思考的问题：文章题目：建议在文章题目中能够直接体现所用选题策略的核心“基于基尼系数”的双目标选题策略。

回应：谢谢专家建议，文章题目已按照专家意见修改。

意见 2：2.引言部分：目前引言部分的内容不够充实，关于各种方法优缺点的述评不够。建议：（1）能略微详细介绍本研究后面模拟研究中比较的几种双目标选题策略，并在后面的模拟研究中简要说明本研究选择这几种方法进行比较的理由。（2）在引言部分增加基尼系数的应用进展，如果没有在教育测量领域的应用，则简要介绍该方法在数据挖掘和决策分类领域的应用，为本研究后面提出的采用这一指标进行新选题策略的合理性做一些铺垫。目前文章中缺少这一部分的文献，请补充。

回应：谢谢专家建议，第二稿增加了第 2 节，详细介绍了本研究中模拟研究部分比较的三种双目标选题策略，在第 4.3 节补充了选择这 3 种方法进行比较的理由。

目前的文献我们没有找到基尼系数在教育测量的应用，在决策分类领域中，CART 算法（Breiman, Friedman, Stone, & Olshen, 1984）是机器学习中比较经典的算法(周志华，2016)，它使用“基尼系数”来选择特征构造决策树获得了较好的分类结果，我们在参考文献中已补充这两篇文献。基于 CART 算法的应用有大量文献，这些文献与本文主旨相关度不高，我们没有引入。

意见 3：选题策略提出部分：（1）目前第二部分的第一大段主要论述了基尼系数和熵之间的

关系。我不清楚这部分的内容是文献中就有的内容，还是你自己推导出来的内容，如果是已有文献中就有的内容，建议放到第一部分叙述且标明出处。另外，这一部分在写作上应突出引入基尼系数的必要性和可行性，还有它与传统方法的联系和区别。在语言表述上建议突出这一点。（2）文中所有公式在出现后都应有对符号和含义的详细介绍，还有一些公式没有标号和文字说明。另外也有一些表述不严谨和符号错误的地方。如（a）第二部分“2 一种基于不确定性度量指标的选题策略”第一段：P(X=xk)=pk, k=1,2,..,n 第三行，n 是不是应该是 K？如果是，建议这里换一个字母，因为后面用 K 表示了属性个数。第 7 行，熵与基尼系数关系的公式有误，请仔细核对，另外请注意此处表述的严谨性，二者是不是相等的关系？其他地方请仔细核对，在此不一一列出。2.1.1 中缺少对引入字母 q 的解释。

回应：谢谢专家提供的行文建议并指出文中公式存在的问题，基尼系数和熵之间的关系推导摘自于网页内容，同时根据高等数学的相关知识也较容易推导。在第二稿中，我们按照您的意见将这部分内容放入第 1 节，并阐述了基尼系数选题策略相对于香农熵选题策略的优势，力求能清楚表达引入基尼系数选题策略的必要性和可行性。

我们仔细核对了文中所有的公式，尽量不再出现此类错误。熵与基尼系数并不是相等的关系，我们在第 1 节作了文字说明。

意见 4：模拟实验研究部分需要进一步详细描述一下内容。 （1）3.2 节下的内容需要重新梳理，按照模拟研究的过程清晰表述，以便于没有做过类似研究的人能够明白模拟的过程。建议按照完整数据生成基于的 Q 矩阵、被试属性掌握模式、数据生成所基于的模型、模型的参数设定、完整作答反应数据的生成、IRT 模型的参数估计以及 CD 模型的参数估计等。并在这部分给出考虑的 CDM 的基本表达式，和每个模型的参数分布。（2）在设计部分，反应数据的生成首先是基于 CD 模型，然后 IRT 的参数是基于生成的反应数据估计得到。我的问题是，是否可以考虑反应数据生成基于 IRT 模型，而 CD 模型的参数估计得到。（3）建议增加选题策略具体实现部分的相关内容。

回应：谢谢专家意见。我们按照您的意见，在第 4.2 节重新梳理了模拟研究的过程，尽量将细节介绍清楚，以便研究者和读者能详细了解整个实验过程。在第二稿的第 2 节和第 3 节阐述了选题策略具体的实现，并在第 4.6 节详细介绍了整个 CAT 的流程，力求能清晰的表达整个实验过程。

双目标 CD-CAT 提供宏观能力评估和微观认知诊断评估，它的题库需涵盖 CD-CAT 和 IRT-CAT 题库中相关信息。CD-CAT 题库与 IRT-CAT 题库除了模型参数不同外，还有非常重要的一点，CD-CAT 题库中的项目需标注项目所测量的属性即 Q 矩阵。目前国际上，双目标 CD-CAT 的研究(Dai, Zhang, & Li, 2016; Kang, Zhang, & Chang, 2017; Wang, Chang, & Douglas, 2012; Wang, Zheng, & Chang, 2014)，都是采用反应数据的生成首先是基于 CD 模型，然后用生成的反应数据估计得到 IRT 的参数。本文借鉴了同类研究范式，采用了相似的设计方法。如果考虑反应数据生成基于 IRT 模型，而 CD 模型的参数估计得到。这种做法，除估计 CD 模型的参数外，还必须估计每个项目的属性向量即 Q 矩阵，而 Q 矩阵的估计本身就是一项十分复杂的工作，目前国内外还没有很理想的 Q 矩阵估计方法。鉴此，本研究还是沿用了国外同类研究的做法，即“反应数据的生成首先是基于 CD 模型，然后 IRT 的参数是基于生成的反应数据估计得到”。当然，我们觉得专家提出的想法“是否可以考虑反应数据生成基于 IRT 模型，而 CD 模型的参数估计得到”很有创意，为我们提供了一个不同的研究设计思路，未来值得进一步探讨。对此，在本文的讨论部分我们也对其进行了补充说明。

意见 5：研究结果部分 （1）目前只给出了每种条件下均值的结果，建议补充标准误（SE）的结果；（2）结果描述过于冗长，请围绕新方法的优劣突出重点趋势和特点，略去其它方法的细节描述；（3）目前图的呈现过于繁琐，如果内容的确较多，建议采用表格形式呈现，且呈现最重要的结果，不用详细展示每个条件和指标。另外图的做法也比较粗糙，建议适当美化。

回应：谢谢专家意见，我们在第 5.1 节补充了分类精度指标的标准误，实验结果表明，对于分类精度指标而言，4 种选题策略的标准误都非常小。

根据专家意见，我们重新梳理了第 5 节的内容，力求能更加清晰的阐述新策略的特性，并重新绘制了文中的数据图及精简了数据，力求能更加清晰的展示 4 种选题策略的特性。

意见 6：讨论部分。目前讨论部分过于单薄，建议扩充更多内容，如对方法之间差异原因的分析和思考；新方法在题库使用方面还有可提高之处，有怎样的改进思路等；本研究的局限性等。总体来看，文章上有较多提升空间，建议修改后再审。

回应：谢谢专家意见，我们重新梳理了第 6 节的内容，分析了各方法之间出现差异的可能原因，并说明了新方法的局限性及可以改进的方向。

第二轮

审稿人 1 意见：

意见 1：作者对上一轮的修改意见做了较好的修改和回应，论文经过上一轮的修改后方法和结果部分有了较大的改进和提高。但是前言和讨论部分还有充实的空间。具体修改意见已经在文档中做了批注，请作者参照批注仔细修改。文中首页第二自然段“……：一是题库建设。……”，根据这一段描述和下一段的方法介绍，困难一更接近模型选择和参数估计，而不是更上位和复杂的题库建设。

回应：谢谢专家指正。的确，正如专家所述，题库建设是更上位和复杂的概念。漆书青，戴海崎和丁树良（1998，P320）指出题库建设是一项系统工程，至少包含（1）学科体系与教育目标层级分类理论；（2）题型功能与命题技术理论；（3）题目分析理论；（4）参数等值理论；（5）包含单题与整卷功效关系在内的测验编制理论等。文中首页第二自然段的内容仅涉及题目分析理论。

根据专家意见，我们将其修改为“……：一是建构题库的心理计量学指标，……”，

意见 2：文中第二自然段末尾“有学者（戴步云,张敏强,黎光明,汪新光,胡姗, 2018; 杜宣宣, 2010; Cheng, 2007; Dai, Zhang, Li, 2016; Kang, Zhang, & Chang, 2017; McGlohen & Chang, 2008; Wang, Chang, & Douglas, 2012; Wang, Zheng, & Chang., 2014; Zheng, He, & Gao, 2018）对此开展了相关探索研究。”，本段的引用文献最好是对两个困难的佐证，比如这些研究曾提出或暗示这两个困难。

回应：谢谢专家的意见。原文中只罗列了这些文献，没有阐述他们的主要研究内容，我们根据您的意见进行了修改，在原处先删除了这些文献的引用，然后再根据这些文献的逻辑关系重新进行了梳理，力求传达 Dual-CAT 选题策略技术发展的脉络，具体修改内容请审阅正文第 2 至 4 页红色标注的文字。

意见 3：文中第 2 页第二自然段“针对困难 1，……”，本文的重点在于新的选题策略，这里介绍分离建模没有体现出与主题的直接关联。

回应：谢谢专家建议，选题策略的构造是基于测量模型的，由于 Dual-CAT 既要测量被试能力又要探查被试知识状态，最佳的解决方案是将两者统一在同一模型中，共同对项目的答对概率产生影响，但目前的文献表明，还没有非常适合的模型能较好的解决这一问题，研究者们采用了另一种方式，在一定条件下，建立被试能力与被试知识状态之间的关系，采用分离模型各自建模，可获得比较稳定的估计。使用统一模型还是使用分离建模这两种方式决定了选题策略的构造方法也不同，因此本文在详细介绍双目标 CAT 选题策略之前，需介绍研究中使用的测量模型即介绍分离建模，从而为后面介绍其他选题策略打下基础。

意见 4：原文中第 2 页第三自然段“针对困难 2，……”，建议将两个困难的说法换为两个重点研究主题？分离建模是设计选题策略的基础问题。
回应：谢谢专家建议，已按照您的建议修改。

意见 5：文中第 2 页第三自然段“Chang(2007)、Wang 等(2012, 2014)、Dai 等(2016)，戴步云等(2018)，Kang 等(2017)，Zheng 等(2018)采用组合策略，即直接选取同时适合当前知识状态估计值$\hat{\alpha}$和被试当前潜在特质估计值$\hat{\theta}$的最优项目。”，建议对诸多策略首先进行更清晰的分类，然后分段描述。
回应：谢谢专家建议，我们按照您的建议进行了修改。诸多策略大致可以分为两个大类，一是以影子题库为代表的策略，二是以组合方式为代表的策略。我们对第二个大类再细分了两种方式：一是加法组合策略；二是乘法组合策略。根据 Zheng, He 和 Gao(2018)的研究这两种组合策略没有谁能绝对占优，需要根据组合策略中使用的信息量准则进行选择。我们对文献中的选题策略进行了归类并分段描述了他们各自的特点，具体内容请审阅第 2 至 4 页红色标注的文字。

意见 6：文中第 3 页第二自然段“选题策略是实施 Dual-CAT 的关键技术，……”，这一段可以与上一段交换位置？
回应：谢谢专家建议，已按照您的建议修改。

意见 7：文中第 3 页第四自然段“熵可用于度量随机变量不确定性，熵越大，随机变量的不确定性就越大。……”，在引入熵的概念之前，建议加入对已有策略的评述，总结得到尚未解决清楚的问题，然后提出从熵的角度设计策略可能是理想方法。
回应：谢谢专家建议。我们补充了对已有策略的评述，在第 4 页的红色文字，具体内容如下：

这些选题策略在一定条件下，都有各自的优势，或精度较高但因运算量大选题耗时较多，如 IPA 策略；或精度稍低用时较少，如 ASI 策略；或精度更低但用时少且题库利用率较均匀，如 JSD 策略。另外这些选题策略，还可能存在两种信息量量纲差异较大或为消除量纲差异进行转换带来信息的损失等问题。我们希望开发一种对$\hat{\alpha}$和$\hat{\theta}$量纲比较统一的信息指标，能保证测量和分类精度较高，兼顾题库利用率均匀性且选题耗时较少的新策略。

意见 8：文中第 2 节，对已有策略介绍在逻辑上的先后顺序是否还有更好的安排方式？或者加入一些承接语？
回应：回复：谢谢专家建议。因我们选择的选题策略均具有一定的代表性，我们补充了选取这些选题策略的原因，具体内容在第 2 节红色文字，例如 ASI 策略是加法组合策略的代表，通过标准化方法消除了两种信息量量纲差异后将转换后的信息量进行线性加权；IPA 策略是乘法组合策略的代表；JSD 策略是题库利用率最均匀且选题耗时最少的策略代表。

希望这些承接语能让已有选题策略的介绍更具有逻辑性。

意见 9：文中第 6.2 节第二自然段，这一段建议分析 Gini 策略在估计精度和用时方面稍差于已有策略的可能原因。

回应：谢谢专家建议。我们分析了 Gini 策略在估计精度和用时方面差于已有策略的可能原因，在 6.2 节补充了这部分内容，具体如下：

在某些条件下（如被试的知识状态由高阶模型生成），Gini 策略的能力估计精度会稍低于 IPA 策略，而此时 Gini 策略的模式判准率会稍高于 IPA 策略，可能的原因是组合策略中能力的信息量和知识状态的信息量共同作用选择下一题，两种信息量在选题过程中互相均衡的结果。Zheng 和 Chang(2016)指出当已知题库参数，公式(3)中的 KL 信息量可以预先计算，缩短了 ASI 策略的选题用时，而 Gini 策略是定义在随机变量的后验概率，必须根据被试的作答反应实时计算，因此选题用时会稍有增加。

意见 10：文中第 6.2 节第三自然段，这一段分析尽量体现 JSD 策略与 Gini 策略的差异和改进方向。

回应：谢谢专家建议。Gini 策略与 JSD 策略的差异主要体现在选题用时和题库利用率指标上，我们分析了差异的原因，并在 6.2 节补充了这部分内容，具体如下：

JSD 策略仅计算基于当前估计值的 KL 距离，运算量小，选题非常快，而 Gini 策略需考虑有限集合和区间范围内后验概率变化，需要求和与积分运算，因此选题耗时会超过 ASI 策略和 JSD 策略。当测验长度较短时，能力估计值和被试知识状态估计值偏离真值较远，基于他们当前估计值的 JSD 策略的选题范围比较宽泛，从而使得题库的利用率会更加均匀；Gini 策略不依赖于能力和知识状态的当前估计值，而依赖于他们的概率分布，选题会更趋集中。

改进方向在下一自然段有阐述：Gini 策略的测验精较高，但其题库利用率不如 JSD 策略。Wang，Chang 和 Huebner(2011)的研究表明限制渐进法（Restrictive Progressive Method：RP）和限制阈值法（Restrictive Threshold Method：RT）能均衡测量精度和项目曝露率，下一步研究拟将 Gini 策略与 RP 和 RT 方法结合，提高 Gini 策略的题库利用均匀性。测量精度和题库利用均匀性是一对相互冲突的指标。使用控制项目曝光技术后，题库利用均匀性会更好，但也会带来测量精度下降的不利影响，如何权衡需要进一步研究。另外，使用控制项目曝光技术后，各选题策略之间的差异是否会消除，也有待进一步研究。当属性个数较多时和题库容量较大时，Gini 策略的选题用时可能会超过用户的期望值，下一步研究拟将 Gini 策略与动态搜索算法(Zheng & Wang, 2017)结合，对其优化以减少选题用时。

审稿人 2 意见：

没有进一步意见了。

回应：谢谢专家审阅。

第三轮

审稿人 1 意见：

意见 1：作者较好的处理了上次稿件中存在的问题。还有一些细节作者需要做出小的修改：仔细核对文章中的公式，尤其是下标，包括公式下面每一个字母和符号解释的准确性。

回应：谢谢专家意见。我们仔细检查了每个公式，并核对了公式中每个字母和符号的解释，保证公式的准确性。

意见 2：经过两次修改，作者为了回复审稿意见，目前文章中有些地方读起来连贯性不太好，建议通读后，对文章的内容和段落中句子的顺序做出必要的调整和修改，以使得文章更加易懂。

回应：谢谢专家意见。我们通读了文章若干遍，修改了一些不太通顺的地方，修改部分已用红色文字标注，供专家审阅，希望修改后的文章能连贯易懂。

意见 3：英文摘要的书写需要进一步提高，建议仔细修改。

回应：谢谢专家意见。英文摘要已请英文较好的老师进行修改和润色，希望能提高英文摘要的书写水平，修改部分已用红色文字标注，供专家审阅。

意见 4：核对参考文献的引用，包括格式以及补充没有引用。 建议修改后发表。

回应：谢谢专家意见。我们仔细检查了正文中文献的引用及格式，正文的引用文献均已在参考文献中，参考文献中所有文献也都在正文中有引用。