《心理学报》审稿意见与作者回应

题目：允许 CAT 题目检查的区块题目袋方法
作者：林喆，陈平，辛涛

第一轮

审稿人 1 意见：
意见 1：simulation 部分交代还不是很清楚，对于纯题目袋（实际为 1 区块）、2 区块、3 区块和无修改 4 种条件应加以描述。特别是题目袋容量的大小是多少？Han(2013)中，将题目袋大小设置为 2、4、6 三种。
回应：参照审稿专家您意见，已在方法部分对 4 种条件进行了更详细的描述。见蓝色部分。

意见 2：.在图 1 的举例中，曾经给出过 5 道题目为一个区块的例子。但在后面的仿真中，30 道题目最多被分为 3 个区块。是否可以增加到 6 个区块，每区块 5 道题目，以证明该方法的效果。
回应：感谢审稿专家提出的问题，确实为了研究的严谨性，应当增加 6 个区块的条件。我们对实验设计和结果进行了补充，见结果部分。结果同研究假设一致，即随着区块数的增加，估计精度提高。区块题目袋方法优于题目袋方法

意见 3：本文方法结合了区块和题目袋，实际上题目袋容量与区块中题目数之比，可能是影响估计精度的重要因素，目前对该因素的考察还不是很充分。

回应：感谢审稿专家提出的问题。如专家所说，这确实是一个重要的影响因素。但本研究通过限定题目袋容量的设计来分析区块的作用，并比较区块题目袋与题目袋方法。所以没有考虑容量和区块的交互效应。因为 Han（2013）的研究已经证实了容量越小题目袋方法的估计精度越高，所以本研究未重复其研究。

专家可能希望提出一个综合性的指标，能够综合考虑题目袋容量，区块数对估计精度的影响。经过我们的认真思考，认为确实有必要增加这样一个影响指标，使之能够更好地应用该方法。事实上，专家建议的影响因素本质上就是区块和题目袋容量对估计精度的影响，这与 Han 的研究和我们的研究假设相一致。所以我们认为在研究假设成立的前提下再来探讨影响指标的建立更好。因此我们在讨论中增加了对影响指标的讨论，见讨论红色部分。

基于研究结果，我们假定估计精度会随着题目袋容量和区块数的变化而变化。那么在定长的终止规则下，这个指标可以表示成关于区块数，题目袋容量的一个函数。如果令题目数为 T（常量），题目袋容量为 P，区块数为 B，那么专家建议的指标 Y=P/（T/B）=PB/T。但由于 P 和 B 对估计精度的影响相反，所以相乘后估计精度并不随这 Y 的减少而增加。

根据专家的建议和我们研究的结论，我们提出在定长终止规则下的一个相对题目袋容量比，Y=100P/TB，该指标意义就是平均每个区块中题目袋容量与总题目数的比。根据我们的研究结论，可以推断出这个指标越小，说明题目袋容量越小，区块数越大，那么估计精度也会越好。 该指标只能为测验开发者提供一定的实践意义，在实践中找到最优的值，也就是最优的设置。在讨论中有进一步说明。另外，在变长的终止规则下，由于 T 变成了一个变量，不同被试的 T 不同，所以该指标仅在定长终止规则下讨论。

意见 4：文章使用 5 种指标来评价能力估计的精确程度，建议把 5 个指标的定义和判断标准加以简单描述，以利于对该话题不是很熟悉的读者阅读。
回应：参照审稿专家的意见，已对评价指标的定义和计算方法列出，并对评价指标的判断标准加以描述，详见方法部分 2.1 最后部分。

意见 5：文中还存在丢字、落字的现象，图 2 和图 3 中纵坐标的名称不一致，请检查。
回应：参照审稿专家您意见，图 2 和图 3 在增加新条件后重新制作，坐标一致，对于文中的丢字，落字现象，实属抱歉，已通读全文进行一一校对，并请同行做了批判性阅读。

意见 6：Han(2013)的文章中还考虑了 IP 使用次数和题目回看次数，对于本研究是否有价值，请作者自行决定。
回应：感谢审稿专家提出的宝贵意见。 这里之所以没有考虑 IP 使用次数和回看次数，主要的原因是觉得这个指标实质上并无有效的价值，因为采用的是模拟研究，对于题目袋的使用次数完全可以根据概率和给定的题目袋容量，区块数进行推断。所以我们认为这个指标的价值不大。

从 Han 的研究看，IP 的使用次数和回看次数的统计也基本符合预期，低能力被试在 IP 容量大的情境下 IP 使用率要大于 IP 容量小的，因为被试从一开始就会遇到很多难题，容量越大自然使用的越多。但回看次数没有差异，因为遇到难题就要回看，在正常作答下，被试遇到难题的概率几乎只跟能力有关，不会因为题目袋容量变大而变化，毕竟没有改变 CAT 的核心程序。随着能力提高，使用次数和回看次数呈下降趋势，因为能力高的被试由于遇到难题的概率更低，但还是容量越大的使用次数越大。回看次数没有差异。 当然，当能力极高的时候，由于几乎不用题目袋，所以使用次数和回看次数极少，也就不受容量的影响了。Han 之所以呈现这两个指标一方面是想使其结果丰满；另一方面是想对题目袋方法产生正向偏差作出一定的解释。而本文充分地考虑到了题目袋方法产生的正向偏差，并通过区块的设置，对其进行了校正。

因此，我们认为在模拟情境下考察这两个指标的价值不大，如果在实际情境中，被试在难度的判断和测验无关因素的影响下，可能在不同题目袋容量和区块数下会有交互效应的出现。这时该两个指标可能会有一定的价值。

审稿人 2 意见：
意见 1：研究议题的提出有待明确和扩充，即为什么要把两个方法进行结合，是逻辑上的原因，而不是根据研究结果再回头提出来的原因。
回应：感谢审稿专家的意见，原文在问题提出中从前人的研究结果肯定了这两种方法的有效性，但随后发现连续区块方法和题目袋方法存在各自的局限与不足，并且给出了这种不足会产生的不良影响。随后给出了充分的理由表明如果两者结合可以相互弥补各自的缺陷，成为一种更好的方法。因此研究假设区块题目袋方法能消除题目袋的方法产生的不良问题，更好地实现允许题目检查的 CAT。我们认为研究的原因确实符合逻辑，并非通过研究的结果来写问题提出，也不是根据结果将两者进行结合。可能在表述上并不清晰，我们对问题提出部分进行了一些修改，烦请专家再细看斟酌，由于内容变化不大，并未进行标注。

意见 2：本文限定了题目袋容量为 6，但未给出原因，也未进行多个变量水平的探讨。而作者在正文中多次强调题目袋容量的设定对结果的影响的重要性。

回应：感谢审稿专家提的意见，本文限定题目袋容量为 6 主要出于 3 点考虑：A、 在 Han（2013）的研究中设置的最大的题目袋容量是 6, 他的研究发现大容量的题目袋会产生大的正偏估计，因此，本研究选取了最大的容量设置。目的就是比较新方法与题目袋方法在容量较大时候的差异。B、为了能固定题目袋总容量，保证题目袋平均分配到各个区块中，所以本研究设定为 6，因为 6 是 2，3，6 的倍数。C、6 个题目袋占了总题目数的 20%，远远超过了一般学生需要修改题目的数量，研究表明一般平均只需要修改 5%左右题目，在这样比真实情境更宽松的模拟条件下进行，更有助于该方法在实践中的应用。我们已在方法部分对为什么设置为 6 容量进行阐述。见 2 蓝色部分。

意见 3：误作答概率如何等同于被试在主观上判断该题为难题的概率，这是一个即需要逻辑解释，也需要数据支撑的。这是本研究的前提假设，而该假设又无法进行检测和操控的（譬如“潜质单维性假设”在实际测验中是可以检测和控制的），因此本文的后续研究犹如一个建构在不稳定且无法人为操控的根基上的房屋。

而 Han（2013）的判断逻辑是，实证数据支持或校准过了，“当题目难度大于能力值 0.5 个单位时”≈“被试有 70%的概率判断为难”。

审稿人认为目前还没有证据表明被试有 50%概率答错（客观值，与认知无关）等同于被试有 50%概率认为该题为难题（主观值，与认知有关）。这两个概率应该不在同一量尺上，因此该假设缺少一个实证数据对量尺的校准过程或研究。

回应：感谢审稿专家对我们研究的模拟过程提出的问题。我们认为 Han 的模拟方式与本文的方式本质上并没有太大差别。他基于的是难度值 b，根据 IRT 模型不难发现，难度值高的被试作答正确率也同样降低。

审稿专家质疑的是 Han 的研究有实证的基础，其结果更符合实际。我们认为这个推论有待商榷，首先，无论哪种模拟方式，对真实情境中题目难度的判断都无法精确的模拟。Han 基于的 Vispoel 的研究也只是一个实证研究的结果，无法一般化。第二，我们认为只考虑难度值 b 是不合理不全面的，区分度同样会影响被试的作答，如果采用的不是极大信息量的选题策略而是 A 分层法选题策略，那么前期被试得到的低区分度题目，用 Han 的方法就会出现问题，因为不同能力被试答对该题目的正确率相似，并不是高能力被试就有极高的把握答对，那么比较 b 值和能力值差在这就不太适用。

所以我们采用了我们自己的模拟办法，我们假设了被试对作答正确率低的题目感觉更难，因为这种假设在很大程度上符合实际情况，所以模拟研究的方式就是通过作答正确率作为判断题目难度的指标。见 2.2 模拟策略 1 中蓝色部分。当然真实情境中被试判断题目难度肯定更加复杂，我们只能采用一种概率较大，更符合实际的方式去模拟。所以我们在讨论中也指出了模拟研究的局限性。 如果非要从量尺的角度去探讨题目难度判断和作答正确率的关系，两者确实存在一定的差别，也可能并非完全的线性关系，但目前，我们没有想到有比正确率更好的办法去模拟被试判断题目的难易。

另外，判断难度的模拟对结果有一定影响，但并不是决定性的。一方面，我们与 Han 的模拟方式有着相似的逻辑。另一方面，虽然不同的模拟方式会影响被试对题目修改的决策，影响题目放入题目袋的行为，但我们认为差别应该不大。区块的作用仍然会体现出来，因为区块的作用是让之前题目袋中的题目提供选题信息，也就是说我们的研究假设无论在哪种模拟条件下，应该不会有太大的变化，因为区块题目袋方法比题目袋方法有更合理的题目袋设置。如果审稿专家仍然有困惑，我们可以之后做个模拟研究，验证两种模拟方式下的结果是否有显著差异。

意见 4：整个模拟研究的作答策略描述过于简单，需要扩充，这是他人重复您研究的必要

条件之一。比如：
1.被试是如何判断题目“为难”？是错误作答概率与随机数比较？
2.一旦判断“为难”就意味着被试一定要把该题目放入题目袋？判断和决策应该是两个认知过程 3.题目袋满后，被试依据什么去挑选正确率最高的题目？这是个主观判断问题。即作者假设被试不仅能判断出哪些题目难，哪些题目容易，而且还能对它们进行主观排序
4.整个过程并没有体现出本文所探讨的“以便之后修改”。

回应：感谢审稿专家提出的建议。本文采用的模拟研究是基于跟实际相符的假设下进行的，关于 1-3 条，我们在方法部分进行了扩充修改。请见 2.2 模拟策略 1 的红色字。由于模拟研究是基于随机抽样和概率模型，确实无法像审稿专家建议的那样将复杂的主观判断和决策纳入到模拟中，因为缺乏实证的基础和模拟方法。我们只能基于先前的假设，在较大的概率程度上实现对真实情境的模拟。
关于第 4 条，这是这个方法的优点，区块题目袋方法给被试提供了一种更符合纸笔测验作答的策略，被试在作答过程中可以对区块中的题目进行任意修改。由于这种独立与选题和临时能力估计，因此，在模拟过程中不需要对被试的作答修改进行模拟，因为该方法只需要这些题目离开题目袋时的最终答案。 所以这个方法几乎不影响对 CAT 的核心过程，更有利于它的推广。

意见 5：审稿人认为本文提出的“类似 Wainer 策略”与 Wainer 策略的本质（认知过程）存在差异，因此命名需要修改。此外，该策略是否能被称为一种“作弊或取巧策略”还有待作者进行确认，因为低能力者在正常作答情况下，也会出现与该策略类似的效果，比如前 6 题该被试均有很高概率答错，也就是作者所假设的有很高概率会把该题目判断“为难”，进而扔到题目袋里。
回应：感谢审稿专家提出的问题。这个命名是 Han（2013）研究中提出的在题目袋方法下被试可能采用的作弊策略。因此本研究为了对照，也进行了这样的命名和模拟。在定义是否为作弊策略时主要考虑的是被试是否有意去操纵 CAT 程序。所以很显然类似 Wainer 策略确实是一种作弊策略。

但如审稿专家所说，低能力被试正常作答也会出现这种类似 Wainer 作答策略，当低能力被试把初始的几个题目都判断为难，确实会出现跟类似 Wainer 作答策略相一致的情况。但这是无意的，个别的，巧合的。在整个模拟中只是极少部分被试的无意行为，显然这种情况下不能认为他们是作弊。同时不能就此否认类似 Wainer 策略不是一种作弊策略。概念上不存混淆的问题。

但在策略 1 模拟过程中确实存在这样的特例使低能力被试作答与策略 2 一致。所以为了减少“作弊”一词对读者造成误解，文章将采用类似 Wainer 作答策略来表述策略 2，并对为什么这么表述作了解释，见 2.3 中的蓝色部分。

专家的提出的问题也说明了这种类似 Wainer 策略在区块题目袋方法中的重要性，是值得深入研究一种作答策略。

第二轮

审稿人 1 意见：
意见 1：请重新绘制图 2 和图 3，确保在黑白打印稿中，能较为清晰的看出图所表现的数据。

回应：感谢审稿专家提出的意见，我们已重新绘制图 2 和图 3，图 2 和图 3 中的 CMAE 的图采用了 2 种不同的线条（虚线和实线）和 5 种易区分的图点。CBIAS 图由于数据本身差异微小，并非制图的问题。

审稿人 2 意见：

意见 1：作者较好地回答了之前的问题。仅从结果上看，新方法的效果确实比连续区块方法好，但根据作者的结果可以看到，实际上作者调整了几个变量后，各个结果之间的差异实在是太小了。在实际应用中，真的有必要为了这么一点点差异而采用这些更为复杂的方法吗？

回应：感谢审稿专家提出的意见，出于以下的考虑，我们认为新方法是有必要的。

根据我们的研究结果已经证实，在模拟的 2 种极端条件下，采用区块题目袋的方法并不会显著降低被试能力估计的精度，而且区块题目袋方法要比纯题目袋方法要好，尤其是当被试采用类似 Wainer 的作弊策略时，区块题目袋方法在所有的能力水平上都要显著优于题目袋方法。 就算只有 2 个区块，依然能很有效地提高估计精度。另一方面，加入适量的区块能够避免题目袋容量过大引起的问题，如果 CAT 的测验长度和题目袋容量进一步增大，我们认为区块题目袋方法会更优异。因为区块能够避免题目袋容量过大带来的诸多问题。

此外，之所以模拟结果差异较小，一方面由于模拟题库是完美题库，每个被试每次选择都能选出信息量很大的题目，因此两种方法均能得到一个较为准确的能力估计值。如果在实际情境中，我们认为差异可能会进一步拉大，题目袋容量过大的影响也会更突出，因此设置区块来减少上述的问题是必须的。

意见 2：GRE 已经不采用 CAT 了，请作者修改文中相关内容。

回应：已对引言部分进行了修改，请见红色部分