# 《心理学报》审稿意见与作者回应

题目：结合 a 分层的兼具项目曝光和广义测验重叠率控制的选题策略
作者：郭磊 王卓然 王丰 边玉芳

## 第一轮

审稿人 1 意见：本研究把兼具同时控制项目曝光率和测验重叠率功能的 SHTOR 法和能够提高题库使用率的 a 分层簇(str-a、str-b、str-c)法相结合，提出了 3 种新的选题策略，以期能够在控制项目曝光率和测验重叠率的同时提高题库的使用效率。根据实验一的结果，发现新的 3 种选题策略较 SHTOR 法能有效提高题库的使用率且测验精度也较高，可以说新模型达到了实验预期。

但该研究仍存在一些不足，关于文章的详细审改建议，请参见文章审改稿，仅供参考：

意见 1：新选题策略也的提出并不明确。在分层法中，按 b 分层、按内容域分层都是有其原因的(作者在文章中也提到了)，但是在实验设计时并未牵扯这些原因，比如 a 参数与 b 参数的相关性并未涉及、内容域的划分比例也未涉及，而这些问题都会影响实验结果。虽然研究提出了 SHTOR-A，SHTOR-B 和 SHTOR-C 三个方法，但并未展现出 SHTOR-B 和 SHTOR-C 的优势。此外，也建议作者修改新选题策略的名称，至少 A/B 应该换成 a/b，因为它们分别对应了 a 参数和 b 参数，而不是 A 参数和 B 参数。

回应：感谢审稿人给出的意见。我们已将 A/B 换成了 a/b，并在实验设计中补充了两个变量：一个是区分度和难度的相关水平，一个是测验考察内容领域的不同比例，请见正文 3.2 部分。并且在结果部分以及讨论部分给出了 SHGT_b 和 SHGT_c 方法的优越性（将 SHTOR 法改成 SHGT 法的理由请见第二位审稿人的意见）。

意见 2：正文书写格式不规范。如序号标准问题、三线表绘制问题等。

回应：感谢审稿人给出的意见。我们根据在心理学报上发表的同类文章修改了三线表（见表 1 至表 4），并且将序号进行了调整和更改（见正文 2.1 部分）。

意见 3：实验二设计存在问题。作者自己给自己挖了一个坑，通过实验二的结果得出了一个暂时无法解决的“对极端能力值出现偏差估计的问题”，且“无效的”补充实验又占据了较大的篇幅。建议作者重新设计实验二，要么寻找到有效解决“对极端能力值出现偏差估计的问题”

的方法，要么直接在实验二中探讨另外的问题，比如新选题策略在不定长终止规则中的表现。

回应：感谢审稿人给出的意见。实验二确实存在您说的这些问题，因此，我们根据您给出的修改建议第一条，增加了两个实验条件，并且删除了一稿中的实验二。

意见 4：调整正文的逻辑顺序，增加文章的逻辑流畅性。比如，在实验一后就给出实验一的结果和结论，而不是等实验二做完后再给出。

回应：感谢审稿人给出的意见。我们已对此问题进行了修改。请见正文“4 研究结果”部分。

意见 5：本研究创新性有限，属于“鸡尾酒研究”。

回应：感谢审稿人给出的意见。本研究根据您给出的修改建议第一条，详细探讨了三种新的选题方法在不同的区分度和难度相关水平以及不同的测验考察内容比例下的表现，得出了理想的实验结果（见正文第四部分），从而验证了 SHGT_b 和 SHGT_c 方法的优越性。从实验结果可以看出，不论实验条件如何改变，题库的使用率均高达 94%以上，这一点不论从理论还是实际应用角度来说，都有很大的贡献。因为建设题库需要投入大量资金，张华华教授曾说到美国考试教育机构每出一道题目，需要消耗 600-1000 美金，因此，在保证被试能力估计精度的同时，能够将题库的使用率提高近 1/3，这种贡献不可忽视。并且将 a 分层思想与 SHGT 法相结合的另一个好处是可以解决传统的 SHGT 法对项目曝光率过度控制的问题，这也是对 SHGT 法弊端的一种改进。

意见 6：不定长 CAT 是什么结果呢？

回应：感谢审稿人给出的意见。考虑到一篇文章，一个研究所关注的点不能太多太散，因此，本文采用大部分研究所采取的方式，即定长 CAT 来对不同的选题策略进行考察，是不失一般性的做法。并且 CAT 在大部分的大型测试中，例如 GRE、ASVAB、GMAT、美国护士资格考试等均是采用定长 CAT 施测的。当然，您给出的意见非常正确，新的方法在不定长 CAT 中表现会是怎样的，并且和定长 CAT 之间的差距有多大，这些都是非常有意义的研究。因此在讨论部分，我们也提到了该观点（见正文最后一部分），可以作为今后研究的一个方向。

意见 7：本研究没有探讨新的选题策略在不定长 CAT 中的表现是较为严重的不足。因为定长 CAT 的实际应用有限，建议作者增加终止策略为不定长的实验。

回应：理由同上。

意见 8：建议作者先参考审稿人意见对文章进行修改，尝试进一步增加创新性。审稿意见：大修后再审。

回应：感谢审稿人给出的意见。我们已对本文进行了大面积的修改。请看正文红色字体部分。

审稿人 2 意见：本研究透過 a 分层， 以提高 SHTOR 的題庫使用率， 研究方法設計嚴謹，研究結果具實務參考價值， 雖然如此， 作者對控管測验重疊率的相關文獻， 似乎尚未能完全掌握， 相較 SHTOR， 已有學者提出更有效率的測验重疊率控管法 (例如：Chen, SY (2010), APM) ，作者宜對這些方法有所說明，其他修改建議如下：

意见 1：SHTOR_A , SHTOR_B, and SHTOR_C 三方法間的優劣比較，作者應有所說明，並提供實務應用上的建議

回应：感谢审稿人给出的意见。正如第一位审稿人的修改建议第一条所述，我们在实验中增加了区分度和难度的相关性和测验考察内容不同比例两个条件，区分出了 SHGT、SHGT_a、SHGT_b 和 SHGT_c 的优劣。第五部分第三段给出了实际应用的建议。

意见 2：评价指标 测验重疊率 T 和 卡方值， 二者具函數關係， 無須重複使用。

回应：感谢审稿人给出的意见。我们已将卡方值删除。

意见 3：作者強調 “四种方法均会低估低能力被试的能力水平（Bias 值在-3 到-1.5 区间内为负值），高估高能力被试的能力水平（Bias 值在 2 到 3 区间内为正值）”， 根據 Bias 的定義， 此論述似乎不正確。

回应：感谢审稿人给出的意见。我们已将 BIAS 改成 RMSE，和 Chen（2010）的做法一致。特此说明：由于本人的粗心，将原来 Bias 所下的结论写反了，但是实验结果是真实可信的。本人对该错误进行了深刻反省，并对结果进行了修正。再次感谢审稿专家指出的问题。

意见 4：“四种方法对中等能力水平被试的估计较为精确，对处于能力量尺两端的被试的能力估计精度较差。”作者雖然進行了補充研究， 結果相似， 作者宜對此結果提供更合理的

解釋 。

回应：感谢审稿人给出的意见。我们不是要回避“对处于能力量尺两端的被试的能力估计精度较差”这一问题，因为该问题是普遍存在与 IRT 及 CAT 领域的，很多研究都得出了类似的结论，更有学者采用加权极大似然估计（Warm, 1989, Psychometrika）以及项目加权似然估计（Tao, Shi, & Chang, 2012, JEBS）来提高被试能力估计精度，特别是对处于能力量尺两端的被试的能力精度，但依然存在“U 型”曲线。因此，我们删除了一稿中的实验二，并根据第一位审稿人的修改建议第一条，增加了两个实验条件。

意见 5：一般來說， 曝光控管愈嚴格， 能力估计的精度愈差， 因此， “当 $r_{max}$ 或 $T_{max}$ 较大时（0.3），四种方法对被试能力估计的精度稍差一些” (P14) ，此論述似乎不太合乎邏輯。

回应：感谢审稿人给出的意见。由于本人粗心，将结论写反了，但是实验结果是真实可信的。本人对该错误进行了深刻反省，并对结果进行了修正。再次感谢审稿专家指出的问题。

意见 6：文獻格式不符合 APA， 此外， 符號前後衝突， 例如：P5 ，i 为题目 vs. 对第 i 个被试 ；k ,...

回应：感谢审稿人给出的意见。我们已对正文的所有符号进行了修正。并对参考文献进行了认真地修改。

第二轮

审稿人意见：本文为修改稿。经过修改，作者将原稿的 SHTOR 法改成修改稿的 SHGT 法，SHGT 法比 SHTOR 法更具有优势。此外，删除了原稿实验二并增加了新的实验。修改稿整体质量优于原稿，但修改稿仍存在一些不足，关于文章的详细审改建议，请参见文章审改稿，仅供参考：

意见 1：正文书写格式依旧不规范。如序号标注问题、三线表绘制问题、字体设置问题、用词不准确，存在病句、赘述等问题。

回应：感谢审稿人给出的意见。我们已通读全文，并按照审稿人的批注建议改正了病句和赘述，以及三线表绘制等问题，见正文高亮部分。

意见 2：修改稿最大的问题在于作者对 c-STR 的理解存在偏差，作者对内容域的操作方法应该是 Kingsbury &Zara (1989)提出的 constrained CAT(C-CAT)法，并不是 c-STR。建议作者建议作者再次阅读 Yi & Chang(2003)的原文或国内关于分层选题策略的相关论文以确定自己的操作方法是否正确，本审稿人认为国内以下两篇文章或许对作者准确理解 c-STR 有所帮助：詹沛达，王立君，杨卫敏.(2013). 引入内容平衡的最大信息量组块分层策略, 江西师范大学学报(自然科学版). 或 程小杨, 丁树良. (2011)子题库量不平衡的按 a 分层选题策略, 江西师范大学学报(自然科学版)。

回应：感谢审稿人给出的意见。由于本人写作上的问题，没有交代清楚，导致审稿人认为我们对 c-STR 的操作是不正确的。我们使用的正是 Yi & Chang(2003)所提出的 c-STR 来对题库进行分层的，而对题库的分层并不能决定实际测验的内容比例。具体操作步骤已在本文“3.1”和“3.2”部分给予的详细说明和更正。本研究固定题库的内容领域数量为 3 个，并且固定每个领域的题目数量均为 120 题，在所有实验条件下均不变。而在实际的测验中，施测者往往是根据不同的测验目的分配不同的测验内容比例，从这个角度来说，探讨测验的内容比例要比探讨题库的内容比例更有价值，因此本研究操作的是测验的内容比例，而固定了题库的内容比例。c-STR 的具体编程是按照本文“2.2.3”所介绍的方法进行的。在这里给审稿人带来的不便，本人致以诚挚的歉意。