《心理学报》审稿意见与作者回应

题目：让自适应测验更知人善选——基于推荐系统的选题策略
作者：王璞珏；刘红云

第一轮

审稿人 1 意见：本研究将推荐系统算法引入 CAT 中，提出了 4 种新的选题方法，从选题视角上，具有一定创新性，但该尝试也面临了许多挑战。通读完全文，我有一些疑惑。

意见 1：CAT 中选题策略有很多种类型，对应不同的选题机制。请问作者选择和两种分层策略比较的原因是什么？

回应：谢谢审稿专家，针对您的疑问，经过深入思考修改了研究设计，将 MIS-B 分层方法替换为最大信息量选题法（简称 FMI），为了考察不同特点的已有答题者数据对推荐选题策略的影响，具体原因在 2.1“生成第一批数据的传统选题策略”进行了详细介绍。

意见 2：通常来说，在 CAT 中的测量精度和题目曝光是此消彼长的关系，而该研究的结果在两个指标上都要比分层方法更优，我对该结果有些疑惑，请作者从原理或方法上给予解释。

回应：谢谢审稿专家，针对您的疑问，重新设计了实验，同时增加重复次数，新的结果验证了推荐选题策略在测量精度与题目曝光率上的权衡，尤其在 FMI 生成数据的情况下最为明显，而在分层策略生成数据的情况下，随着重复次数增多可以发现，在三种测量精度指标上小于 0.02 的差异属于随机波动的正常范围，分层策略与推荐策略的精度差异很小，更严谨的说应当是在不损失精度的情况下小幅改善了题目曝光。原稿中重复次数较少，对结果的解读和表述不够严谨，现已修改，详见 3.3 和 4.2 两个模拟研究的结果。

意见 3：“……由此可见，从已有方法出发，结合新的思想，形成更优的解决方案，是改进 CAT 选题策略的一种常见且有效的途径”——作者提到的新思想是指什么思想？

回应：谢谢审稿专家，原文中“新思想”指的是改进选题策略时全新的思路和理论，例如本文中基于协同过滤推荐的假设对于基于 IRT 的传统选题方法可以称为新思想。由于引言部分进行了较大调整，这句话已删去。

意见 4：“Chen 等人（2018）……通过模拟研究发现使用推荐系统比随机选择学习材料在不同的奖励函数上都得到了更优的结果”——能否稍加解释结果表达的含义？没有看过 Chen 文章的读者可能不太理解“在不同的奖励函数上都得到了更优的结果”想表达什么。

回应：谢谢审稿专家，原文中奖励函数等概念涉及强化学习和部分统计定理，结论较为复杂，为了便于读者理解，已改为“……通过模拟研究发现使用推荐系统选择学习材料比随机选择以两种统计指标衡量都有更高的效率”，见引言第三段（采用蓝色字体）。

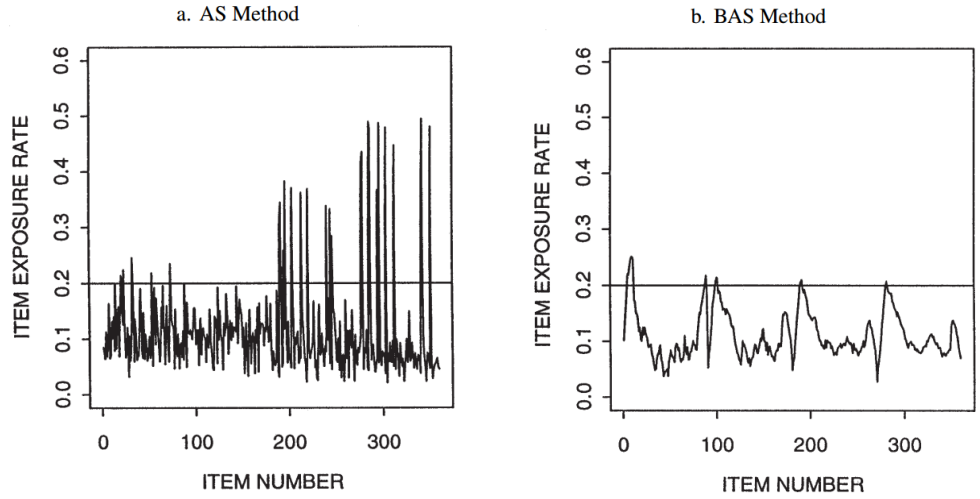
意见 5：“因此已有答题者数据的质量决定了基于推荐系统的选题策略的表现。”——模拟研究中可以控制住该条件，但在实际使用过程中，如何保证该前提要求得到满足？换句话说，用什么方法检验已有答题者数据的质量呢？

回应：谢谢审稿专家，参照您的意见，新的研究设计中选用 FMI 与 BAS 形成对照，生成的答题者数据在精度和曝光的各指标上的特点差异鲜明，推荐选题策略也随之表现出不同趋势。借用这种方式对已有答题者数据的质量进行评价，即通过各指标上的绝对大小，以及在

精度和曝光上侧重不同的特点共同衡量。

意见 6：“BAS 不但继承了 a-分层的优点……又减少了过度曝光和曝光不足的题目数”——根据 Parshall 等的研究指出，a 分层方法仍会出现题目过度曝光的现象。Parshall, C., Harmes, J. C., & Kromrey, J. D. (2000). Item exposure control in computer-adaptive testing: the use of freezing to augment stratification. *Florida Journal of Educational Research*, *40*(1), 28-52.

回应：谢谢审稿专家，已阅读您给出的文献，发现自己原稿的表述确有不清楚之处，a-分层会出现题目过度曝光的问题，而非优点，Chang, Qian, & Ying 在 2001 年提出的 BAS 方法在一定程度上改善了 a-分层的这一问题（图为原文中模拟研究的结果），这也是本研究选择 BAS 方法的原因之一。已修改了原表述，并加入您给出的文献，见 2.1“生成第一批数据的传统选题策略”第二段 （采用蓝色字体）。



Chang, H. H., Qian, J. H., & Ying, Z. L. (2001). a-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*, 333–341.

意见 7：“可以发现本研究是以较为简单的二分方式计算答题者的相似度”——请进一步阐释，什么是二分方式，怎么使用二分方式呢？

回应：谢谢审稿专家，参照您的建议，已经对介绍相似答题者的内容进行了修改，见 2.2“基于协同过滤推荐的新选题策略”第一段（采用蓝色字体）。简而言之，推荐系统中常用向量的余弦值计算相似度，得到的是连续值；对于推荐选题，如果在相同题目上作答相同即判定为相似答题者，不满足条件则不是，得到相似性结果的是 0 或 1，故称为二分方式。

意见 8：“找到相似答题者后，将他们回答的下一道题目与当前答题者未作答题目的交集作为备选题目，从中随机抽取一题作为当前答题者的下一道题目”——为什么是随机抽取？

回应：谢谢审稿专家，选择随机抽取有两个原因：第一，随机化操作是一类常用的可以控制题目曝光率的方法（Georgiadou, Triantafillou, & Economides, 2007），本研究的结果表明两种加入随机抽取的新策略对题目曝光率控制很好。第二，最初设计策略时曾尝试找到备选题目后，将随机抽取（表 1 中 DEBR 和 IEBR）变为匹配当前题目的位置，选择位置最接近的题目推荐（表 1 中标黄的 M-DEBR 和 M-IEBR），预实验的结果表明没有提高精度，在题目曝光率控制上优势也不如随机抽取，故最终选择了随机抽取。

表 1 对比随机选题和精确匹配选题的预实验结果

	MAE	MSE	χ^2	Overlap Rate
BAS	0.374	0.229	14.348	9.09%
DEBR(BAS)	0.374	0.232	10.868	8.14%
M-DEBR(BAS)	0.375	0.227	13.878	9.06%
IEBR(BAS)	0.368	0.221	11.363	8.29%
M-IEBR(BAS)	0.371	0.227	13.935	8.96%

注：括号内为生成已有答题者数据的选题策略。

意见 9：“将当前答题者未作答题目中 b 参数（2PLM）或 θ_{max} 值（3PLM）位于该范围中的题目作为备选题目”——题目参数和 θ_{max} 值有何关系？如何选择呢？

回应：谢谢审稿专家，原稿中 θ_{max} 值是指使信息量达到最大时的能力参数值，MIS-B 在选题时是寻找 θ_{max} 值与当前能力估计值最接近的题目，使该题目对当前答题者的信息量最大。现已将 MIS-B 替换，不再涉及这一问题。

意见 10：一道题目会有不同被试来做，这道题目对于某些人会出现在 CAT 前期，有些会出现在后期。而 CAT 前期对被试能力估计的精度是不高的。基于目前的操作，会存在把剩余题库中大部分题目都作为备选题的可能性，基于此再随机抽取，这种选题法会更好吗？其次，也会存在缩短或增大最小值和最大值间距的可能性，这些因素都会影响估计精度。

回应：谢谢审稿专家，您的第一个疑问在审稿人 2 意见 8 的回应中已做出部分说明，考虑出现的位置没有带来明显的精度升高。此外，基于目前的操作，备选题目的范围没有您预计的那么大。以研究二（276 道题的真实题库）为例，使用 FMI 生成的数据，备选范围平均为 13 道题，使用 BAS 生成的数据，备选范围平均为 3 道题。本研究重在将推荐系统的思想引入 CAT 选题，提出的新策略较为简单，确实会出现能力范围的间距不准确的问题，未来可针对精度问题进行更复杂有效的改进，在讨论部分的最后一段中提出了多个提高精度和测验初期选题的改进建议，如第一、二、四、五（采用蓝色字体）。

意见 11：如果找不到相似答题者该怎么办呢？由此，又会带来 2 个现实问题：①实践中，应该有多少已有答题者数据才能够使用本文提出的新方法。②题库中题目数量和已有作答者数量之间应该满足什么要求？（可以是比例，也可以是其他）

回应：谢谢审稿专家，针对您的疑问，仔细检查结果后发现，找不到相似答题者的主要原因是已有答题者数据中题目使用是否均匀，也就是生成该数据的策略是否能有效控制曝光：使用 FMI 生成的第一批数据更容易找不到相似答题者，将其与推荐策略生成的第二批数据合并后便找不到的概率便小于万分之一，而使用 BAS 的第一批数据发生概率便小于万分之一。现实中推荐系统的用户数会多于项目数几个量级，在 CAT 下则宽松许多。本研究中一批答题者的数量设定为 1000 人，对于两个模拟研究中使用的题库（400 道与 276 道）都满足要求。由于首批答题者数据可以通过模拟生成，答题者数量可自行设定为很大的数量。本研究的结果来看，大致推断答题者人数几倍于题目数，几千人即可，重在初始的选题策略。在讨论部分第四段中讨论了这个问题（采用蓝色字体）。

意见 12：“为避免极端值的影响，计算出他们所有当前能力估计值的中数，将这一结果看作当前答题者的近似能力估计值”——为何是中数呢？

回应：谢谢审稿专家，在最初设计策略时尝试过基于取平均数的策略（表 2 中标黄的 MEAN），预实验结果表明选题精度和题目曝光都差于取中数的策略（表 2 中 MIBR 和

MEBR），分析原因可能是平均数受极端值影响较大，加之每道题筛选出的相似答题者数量不多，尤其在测验初期能力估计不准时，个别能力估计偏差较大的人会影响最终结果，而中数对极端值不太敏感，选题结果更为稳定。

表 2 对比计算中数和平均数选题的预实验结果

	MAE	MSE	χ^2	Overlap Rate
BAS	0.337	0.200	3.070	11.9%
MIBR(BAS)	0.307	0.159	3.163	11.9%
MEBR(BAS)	0.317	0.164	5.047	12.4%
MEAN(BAS)	0.320	0.169	9.059	13.5%

注：括号内为生成已有答题者数据的选题策略。

由于在修改稿中删去了 MIBR 和 MEBR（具体原因见审稿人 2 意见 23 的回应），不再涉及该问题。

意见 13：A 参数通常服从对数正态分布？这些参数的范围为何取这些值？是有什么特殊考虑吗？还是和实践中比较吻合？

回应：谢谢审稿专家，本研究对各参数的分布设定沿用了 Cheng, Patton, & Shao（2015）进行选题策略比较研究时使用的模拟参数：

: $a \sim N(1.2, .25)$, $b \sim N(0, 1)$ and $c \sim N(.25, .02)$. For each item bank, ated randomly, with trait levels extracted from a distribution $N(0, 1)$. 20 and 40 items, were used. The initial trait level, $\hat{\theta}_0$, was selected

Cheng, Y., Patton, J. M., & Shao, C. (2015). a-Stratified Computerized Adaptive Testing in the Presence of Calibration Error. *Educational and Psychological Measurement*, *75*, 260–283.

意见 14：结合本研究的条件，有个问题：假如现在估计新抽取的第 2 号被试，推荐系统是利用了 1000 个已有答题者的数据，还是 1000+1 个答题者数据？

回应：谢谢审稿专家，本研究每次都以一整批答题者的数据进行选题，这样便于衡量第二批答题者的整体选题情况。如果以第二批答题者中每人基于的数据量不同，比较难以进行不同批次的比较。在现实应用中，可以将每一位的作答结果都加入下一位的参考数据中，在本研究中为了便于批与批之间比较，总以 1000 人为单位。已针对您的疑问修改了表述，见 3.1 和 4.1 研究设计（采用蓝色字体）。

意见 15：如果对于需要新估计的被试能力分布和已有答题者的能力分布不同的话，会对本研究的方法产生什么影响吗？

回应：谢谢审稿专家，您提出的疑问确实是非常重要的一个影响因素，限于本文篇幅所限，未能在模拟实验中加入该因素进行探究，现已写在讨论部分倒数第二段本文局限中，作为第二个未来可继续探索的方向（采用蓝色字体）。

意见 16：通常，CAT 中的 a 和 b 参数都是有相关的，请问作者设置无相关的原因是？

回应：谢谢审稿专家，本研究对题目参数相关的设定同样沿用了 Cheng, Patton, & Shao（2015）进行选题策略比较研究时使用的模拟参数：

Wingersky and Lord (1984) and H. H. Chang et al. (2001) have pointed out that in practice, the *a* and *b* parameters of the items are usually correlated. The performance of different item selection rules can vary depending on whether the item banks used have correlated parameters or not (Barrada, Olea, et al., 2009). Thus, two kinds of item banks were generated, one with uncorrelated *a* and *b* parameters and the other with correlated parameters ($r_{ab} = .5$). A total of 10 banks

看到您提出的疑问后，再次查阅了 CAT 相关文献，尤其是使用分层策略的研究，考虑到 BAS 的提出正是考虑到题目参数的相关性，大部分研究主要关注的也是有相关的情况，综合考虑您的意见，去掉了无相关的情况。

意见 17：“研究一共生成四个模拟题库（题目参数见表 1）。每种条件下重复 10 次。”——重复条件偏少。
回应：谢谢审稿专家，已将重复次数修改为 100 次。

意见 18：“上述四个指标的值越小，表示题目曝光控制得越好。”——曝光不足不是越小越好。
回应：谢谢审稿专家，已去掉该表述，并参照使用该指标的文献（He, Diao, & Hauser, 2014）修改了对曝光不足的标准。

意见 19：“因此在题库确定的情况下，如何提高已有答题者数据的质量就显得至关重要。”——这也是实践中特别要注意的一点，但该如何保证呢？或者该如何检验质量呢？
回应：谢谢审稿专家，您的疑问在审稿人 2 意见 5 的回应中做出了部分回答。为了更好体现出已有答题者在质量方面的差异性，便于研究这个问题，新的研究设计中去掉了一种分层方法 MIS-B，换成了重测量精度轻曝光控制的 FMI，与 BAS 生成的数据相比在各项指标上呈现出明显的差异，得以对推荐选题策略的特点有很全面的了解。基于两个研究的结果总结，可以通过各指标上的绝对大小，以及在精度和曝光上的相对侧重趋势衡量已有答题者数据的质量和特点，进而预测推荐选题的表现。

意见 20：“使用推荐选题策略的同时也在产生已有答题者数据，且该数据的质量优于传统选题策略第一次生成的已有答题者数据”——这个结论是从何而得的呢？根据研究一的结果，传统方法的估计精度并不差。
回应：谢谢审稿专家，通过修改实验设计，提高重复次数，新的结果表明推荐选题策略与 BAS 方法在精度方面的差异很小，已修改原有结论，见 3.3 和 4.2 研究结果（采用蓝色字体）。

意见 21：“使用四种推荐选题策略（1）结合自身生成的已有答题者数据，（2）结合 MIS-B 和自身生成的已有答题者数据，对 1000 名新答题者（第三批）进行 CAT 模拟”——没有描述清楚。
（1） 是只考虑了第二批 1000 的数据
（2） 是考虑了第一批和第二批，共 2000 人的数据吗？
本研究设置了两个实验条件，结果有 3 个表，表 4 是什么条件下的结果？
回应：谢谢审稿专家，已对原有设计进行了修改，现仅保留（2）的条件，即合并第一批和第二批数据，共 2000 人，同时修改了表述，精简了表的数量，见 4.1 研究设计和 4.2 研究结果。

意见 22：“正如在估计方法中最大后验估计解决了极大似然法在作答全部正确或错误时无法估计答题者能力的问题”——该举例不恰当，贝叶斯法和 ML 是两种思路，不是说加入了先验信息，贝叶斯就克服了 ML 的缺陷。
回应：谢谢审稿专家，已删去该例子。

意见 23：“不借助任何题目和能力参数即完成了选题”——该句话不恰当。推荐系统在选题时仍然要匹配 b 参数，还是要用到题目和能力参数！
回应：谢谢审稿专家，已删去该表述。深入学习推荐系统并请教相关从业者后，同时发现了原稿中对协同过滤的理解和对选题策略命名有误。为了更清楚地让读者理解协同过滤的原理和推荐选题策略的设计思路，现将基于题目推荐（Item-Based Recommender, IBR）改名为直接的基于答题者推荐（Direct Examinee-Based Recommender, 简称 DEBR），将答题者推荐（Examinee-Based Recommender, EBR）改名为间接的基于答题者推荐（Indirect Examinee-Based Recommender, 简称 IEBR），同时将容易引起混淆的 MIBE 和 MEBR 删去。并在 2.2“基于协同过滤推荐的新选题策略”中解释了底层假设和最终实际操作上的差别，在设计思路上“*找到相似答题者后，可借鉴和协同过滤推荐的底层假设。一种改良假设是：当前答题者可以使用与相似答题者相同的下一道题目，这样便得到一种直接的推荐选题策略，不借助题目参数完成选题*”，而在实际操作中“*上述两种推荐选题策略有可能找不到可推荐的题目，称为选题失败……找不到可推荐题目时便使用当前答题者的能力估计值匹配 b 参数选择下一道题目*”，最后在讨论部分中着重强调了“*推荐选题策略并非完全抛弃已有的选题思想，而是推荐系统和传统选题方法有机结合的产物。无论是直接还是间接基于答题者推荐选题，在题目曝光率控制方面有优秀的表现，很大程度上都得益于加入了随机选题和匹配 b 参数的选题操作*。”

意见 24：“本研究先使用分层策略生成已有答题者数据”——如果是其他方法生成的数据，会对结果有什么影响吗？新方法的表现还会好吗？
回应：谢谢审稿专家，综合考虑您的多条意见后，修改了研究设计，将另一个分层策略 MIS-B 替换为特点不同的最大信息量选题（FMI），生成的数据更注重精度而曝光控制不佳，推荐选题策略的表现确实与 BAS 分层策略下不同，表现出明显的减少未使用题目，大幅改善题目曝光率的趋势，但以损失一定精度为代价，直接的基于答题者推荐（DEBR）损失精度较少，精度高于同等条件下的 BAS，间接的基于答题者推荐（IEBR）对曝光率改善幅度最大，甚至优于 BAS，精度与 BAS 持平，详见 4.2 研究结果。

审稿人 2 意见：文章将基于推荐的选题策略与传统的 CAT 选题策略相结合，这在当前的国内外的 CAT 研究中还十分少见，具有创新性。现有文章语言比较通畅，但整个行文逻辑结构还需要调整，另外，文中存在一些细节需要修改，具体意见如下：
意见 1：文章的逻辑结构需要进行大的调整。第一部分引言，通常在英文文献中引言可以包含所有的背景、文献综述和问题提出部分，作者的引言主要是介绍 CAT，以及推荐系统的现状等等，更类似于背景，但作为背景在文中占比过大，需要删减。如第二页第二段最后“数据挖掘比赛在 2010 年提供……获得当届比赛第三名的好成绩。”这样的事实描述性的东西，只适合作为论文开始的很小一部分出现。要说明推荐系统的优势，应该从方法层面或理论层面更科学化的论述。
回应：谢谢审稿专家，参照您的意见，同时查阅了若干英文文献和《心理学报》上发表的方法创新类的文献，重新组织了引言的结构，精简 CAT 和推荐系统的背景介绍，主要分析了

CAT 的现有问题，着重阐述了推荐系统的优势，从原理层面分析了将推荐系统引入 CAT 选题的重要性和可行性，并在引言结尾处补充了本研究试图解决的问题。

意见 2：第二部分“基于协同过滤推荐的选题策略”，这个部分里面夹杂了文献综述、方法介绍、问题提出、解决方法等等，比较混乱。鉴于引言中没有太多文献综述的部分，且本文是方法类的论文，可用方法介绍作为文献综述。首先介绍其他更多的选题方法，重点强调 BAS 方法。再谈 BAS 的优势（也是作者选择该方法的理由）。BAS 的方法具体介绍可放在方法部分，也可以简化，给出引用。传统 CAT 选题介绍完后，再引入推荐系统的方法，作者需要先归纳总结已有的系统，再重点强调文章所选系统。说明和其它系统比的优势，说明和传统 CAT 选题方法比，推荐系统的优势。总的来说，第二部分的意义在于通过归纳现有研究，突出现有研究方法的先进性，优越性，而非描述现有方法在干什么，怎么做。

回应：谢谢审稿专家的意见，查阅已有研究选题策略的中英文文献后发现，常见的一种结构是第一部分引言主要通过总结现状发现存在的问题，确定改进的方向，第二部分主要阐述现有的方法和新提出的方法的原理，第三部分设定参数进行模拟实验。同时参照您的意见，做出如下修改：在引言部分阐述引入推荐系统改善 CAT 选题的必要性之后，将第二部分重新梳理，主要介绍本研究要用到的几种选题策略，强调了 BAS 和新加入的 FMI 方法的优势，最后介绍新方法的原理和操作过程。

意见 3：模拟研究需要继续补充、修改。第一，作者只考虑了 40 题的终止规则，对于很多 CAT 的研究而言 40 题已经不少了，且单一题长情境无法判断选题方法的变化情况，还是仅仅只在一个情境下更好。第二，模拟研究只重复 10 次是远远不够的，考虑到作者使用的 2PLM 和 3PLM 并不是特别复杂的模型，即使 100 次的重复计算负担也并不大。第三，设定为什么要设定 a 和 b 的相关情境，这实际上和作者选用的选题方法有关，但作者需要在模拟设定部分具体说明原因以及如何模拟相关的，且同时说明为什么只设定了中等相关的情境和无相关的情境比较。

回应：谢谢审稿专家，参照您的意见，在两个研究中增加了 20 道题目的条件，将重复次数提高到 100 次，综合考虑另一位专家的意见，去掉了无相关的情境，保留中等相关的情境（具体原因见审稿人 2 意见 16 的回应），修改见 3.1 和 4.1 研究设计部分（采用蓝色字体）。相关的设定是参考了研究分层方法和做策略比较时常用的模拟参数（如 Barrada, Olea, & Abad, 2010; Cheng, Patton, & Shao, 2015，相关分别为 0.45 和 0.5）。由于本研究中 a 和 b 都服从正态分布，常用的统计软件中（如 R 语言）有生成多元正态数据的函数，可以在设定相关的条件下模拟得到 a 和 b 参数。

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. Applied Psychological Measurement, 34, 438–452.

Cheng, Y., Patton, J. M., & Shao, C. (2015). a-Stratified Computerized Adaptive Testing in the Presence of Calibration Error. *Educational and Psychological Measurement, 75*, 260–283.

意见 4：表 1 可删去，用文字简单描述即可。

回应：谢谢审稿专家，参照您的意见，已删去表 1，改为文字描述，见 3.1 和 4.1 研究设计（采用蓝色字体）。

意见 5：国内的习惯 3PLM 的 c 参数一般就直接称为猜测参数。

回应：谢谢审稿专家，参照您的意见，已改为猜测参数。

意见 6：偏差（bias）（3.2 评价指标）这个指标是没有意义的，由于能力估计使用贝叶斯后验，其估计结果的分布均值为 0，不考虑正负的情况下，差值的平均值都为 0。建议删去该指标。
回应：谢谢审稿专家，参照您的意见，已删去该指标。

意见 7：公式（5）中的曝光率 r_i 和公式（6）曝光率的方差是如何求的，作者或者描述清楚，或者给出引用，以方便不了解的读者重复和追溯。
回应：谢谢审稿专家，参照您的意见，在介绍两个指标的文字后都已加入文献来源方便读者查阅，见 3.2 评价指标（采用蓝色字体）。

意见 8：如文章中提到的，推荐系统的优劣极度依赖前期的作答信息。然而，被试的前期作答信息是怎样的，作者需要说明具体。例如，前期作答中被试能力分布如何？有多少被试？前期被试的作答模式如何？
回应：谢谢审稿专家，参照您的意见，将研究设计部分重新撰写，强调了每一批被试的数量、能力分布等情况，见 3.1 和 4.1 研究设计（采用蓝色字体）。

意见 9：接上一问题，推荐系统的优劣基于前期被试信息的准确性，那么如果被试信息不准确将如何，同样对于所有 CAT 而言早起阶段的选题准确性很大的影响了后面的结果，因此希望作者在讨论部分中加入相关讨论，也期望在未来研究中能作为一个探讨角度。
回应：谢谢审稿专家，参照您的意见，为了考察前期信息准确性对推荐选题的影响，将 MIS-B 分层方法替换为最大信息量选题法（文中简称 FMI），着重考察了在测量准确度不同的答题者数据如何影响推荐的结果。对于 CAT 初始阶段选题的问题，已加入到讨论部分未来可改进的方向，提出了借鉴推荐系统中解决冷启动问题（因信息不足造成推荐困难）的相关技术，见引言最后一段第四点（采用蓝色字体）。

第二轮

审稿人 1 意见：作者对两位审稿人的审稿意见做了相应回答，在论文逻辑和研究设计上有了进一步提升。针对修改稿，审稿人还有以下几点意见供作者参考。
意见 1：P17 第三段介绍 IEBR 时，作者说“IEBR 将统计所有相似答题者答完本题后能力估计值的范围……”。请问作者相似答题者的能力估计值是相似答题者最终的能力值，还是作答到当前这道题目的能力估计值。如果是前者情况，那么相似答题者的能力可能已经发生较大偏移了；若是后者，意味着并非他们的最“真实”能力值，也会增大选题误差。请作者对该问题进行解释和说明。
回应：谢谢审稿专家。文中 IEBR 使用的是作答完当前题目之后的能力估计值，也就是您给出的第二种情况。如您所言，从 IEBR 的实现过程来看，无论使用答完当前题目还是答完所有题目的能力估计值，都可能存在或多或少的误差，这是使用已有答题者数据时较难避免的，也是降低测验重叠率的一个结果，即没有一个人的已有数据可以直接且精准地借用。为了解决这一问题，IEBR 的设计思路是推荐系统与传统选题策略相辅相成，保留传统选题策略中根据当前能力估计值选择下一道题目的思想，以推荐系统中计算用户相似度的协同过滤方法找到可选题目，同时借鉴了不止一种控制题目曝光率的操作。模拟研究的结果初步证实了 IEBR 的整体思路是可行的，随着答题者所做题目的增加，能力估计的精度逐步提高。为了

使表述更加清楚，避免读者产生此类歧义，已将原来的“IEBR 将统计所有相似答题者答完本题后能力估计值的范围……”修改为“IEBR 将统计所有相似答题者答完本题后得到的当前能力估计值的范围……”，见“2.2 基于协同过滤推荐的新选题策略”第二段蓝色字体修改部分。

意见 2：评价指标部分，对于公式中首次出现的符号要给予解释。
回应：谢谢审稿专家。已将所有首次出现的符号都加上解释，见“3.2 评价指标”蓝色字体修改部分。

意见 3：研究二中的题库既有 2 参数题目，也有 3 参数题目，请问作者是如何完成被试能力参数估计的？类似于这种“混合题库”是否有参考的文献支持？
回应：谢谢审稿专家。本研究使用了 R 语言中一个专门用于 CAT 的软件包 catR，被试能力估计使用了 catR 中的 thetaEst 函数，该函数可针对多种 IRT 模型和估计方法完成能力估计，只需将作答过 n 道题目的题目参数设定为一个 n 行 4 列的矩阵格式（默认的 0-1 计分答题模型为 4PLM，若为 2PLM 和 3PLM，将对应不存在的参数设为 0），即可得到该答题者的能力估计值，完成“混合题库”的能力参数估计。具体用法和原理可见参考文献：
Magis, D., & Raîche, G. (2011). catR: An R package for computerized adaptive testing. Applied Psychological Measurement, 35(7), 576-577.
Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. Journal of Statistical Software, 48(8), 1-31.
Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. Journal of Statistical Software, 76(1), 1-19.

意见 4：对 4.2 部分研究结果的表 3 有所疑问。由于作者没有继续迭代第三轮，根据表 3 数据结果来看，同时考虑前两批作答数据的推荐选题策略的估计精度都在变差，是否再多做一轮的估计精度还会变差呢？如果是，那么推荐选题策略可能会存在较大问题。
回应：谢谢审稿专家。详细分析研究二的结果可以发现：由 BAS 生成数据时，两轮迭代后的精度没有明显下降（从 0.253 到 0.262；0.266 到 0.267），100 次重复后三种精度指标取均值后，不超过 0.01 的差异属于正常波动，与此同时卡方值、重叠率和过度曝光都明显下降，表现出比 BAS 更好的题目曝光度控制能力；由 FMI 生成数据时，首要问题是题库中存在大量从未使用过的题目，此时取得的高精度实则损害了题库安全，两种推荐选题策略在两轮迭代中使用的数据都含有 FMI 生成的曝光控制不良的首批数据，最终都较好地完成了启用整个题库，优化题目曝光率的首要目标，在此过程中损失一定的精度是 CAT 选题中常见的一种权衡，也从另一个角度证实了推荐选题策略可以在两轮迭代后有效弥补 FMI 在题目曝光方面的不足，而且 DEBR 的精度始终高于 BAS 方法，IEBR 也一直远高于随机选题，没有过度损害精度。更重要的，BAS 条件下的结果表明：当已有答题者数据不存在极端的曝光不均时，推荐选题策略优化曝光率控制便不再以损失精度为代价，而 FMI 下第二轮迭代后曝光不均的问题均已良好解决，可推知之后若干轮迭代的结果也将是保持精度并优化曝光控制。综上，推荐选题策略的最终目标是找到现有条件下精度和曝光控制较好的平衡点，精度降低是某些极端情况下不可避免的权衡之举。在本研究的首轮文稿中，曾有真实题库和分层策略下，多种推荐策略先结合自身生成数据再结合所有历史数据的三轮模拟研究，选题精度没有明显的下降。

意见 5：P23 第二段中“这使推荐选题策略更少受到 IRT 的前提假设（如单维性或条件独立性）不满足所产生的影响”。该推论可能存在问题。因为不论是 DEBR 还是 IEBR，都要与已有

作答者产生关系，而已有作答者的数据追溯到源头，还是要与 IRT 有关系。

回应：谢谢审稿专家。如您所言，现阶段生成第一批数据确实需要一个基于 IRT 的选题策略，主要因为现有基于 IRT 的选题策略研究已经非常成熟，特点明确，便于研究设计的操控和比较。本文主旨在于探索用推荐系统进行 CAT 选题的可能性，依托的协同过滤推荐技术需要先有第一批数据，但就 DEBR 和 IEBR 方法本身来讲，这批数据可以是基于 IRT 选题策略生成的，也可以不是基于 IRT 选题策略生成的，本文的确没有涉及到不借用传统策略生成第一批数据的其它完全基于推荐系统的实现方法。为了使表述更加准确，同时也不失 DEBR 和 IEBR 方法本身的可拓展性，已将原句修改为“随着研究不断深入，尤其是推荐系统的更多引入，可能会在生成首批数据或预防选题失败等方面逐渐摆脱对传统选题策略的依赖……”，详见讨论第二段蓝色字体修改部分。

意见 6：推荐选题策略的优势是多利用了先有作答者的数据，但根据模拟结果，其精度和曝光程度还是介于 FMI（代表精度最高）和 BAS（代表曝光控制较好）之间，因此，题目中的“更”可能不太适合。可以考虑：基于推荐系统的自适应测验选题策略，或者其他。

回应：谢谢审稿专家。您的意见十分重要，促使我们对模拟结果和文章立意进行了更深入地分析和思考。综合两个模拟研究的结果可以发现，对于使用他人数据的推荐选题策略，最首要的影响因素和评价对照都是推荐策略使用的已有数据自身的特点和质量。在几乎所有条件下两种新策略都比它使用的数据曝光控制更好，且在所有 BAS 条件下新策略的精度都没有实质的下降，考虑到 FMI 是过度牺牲曝光控制才达到了每个条件下精度上限，而新策略在设计时着重加强了控制曝光方面，最终结果是新策略达到了设计预期并做到了尽可能少的精度损失，很少低于 BAS。大体上看新策略的两大类指标位于 FMI 和 BAS 之间，实际可以认为：推荐选题策略在达到与 BAS 同样好的曝光控制的同时，精度更高更接近 FMI，这是文章标题中“更知人善选”的本意和具体体现。另一方面，本研究发现了推荐系统和 CAT 的共通性，推荐选题策略的提出建立了一个全新的灵活的框架，未来还有使适应性测验更加精准和智能的提升空间，此为题目更深一层的含义。

……

审稿人 2 意见：本研究在上一次的修改基础上做了有效的补充和修改，但论文整体仍显出重点不突出的问题。另外研究内容仍需进一步补充。具体建议如下：

意见 1：文章篇幅太长，需要进行大的删减，作者所有的文字应该是围绕这研究论文的目的服务的，而非阐述事实。以第二段为例，“作为重要的研究方向，推荐系统时常出现在计算机学术会议和数据挖掘比赛中”这样一句话类似于新闻描述，之后还附带了三篇文献的引用。这之后的一段“如今，各大在线购物、社交、娱乐网站和公司能为用户提供精准贴心的服务，无不有一套强大的推荐系统作技术支持……”这同样类似于现状描述，实际上作者的目的是想说明推荐系统有作用优势。“当前推荐系统的算法可以根据用户的需求给出精准的匹配，这在商业、娱乐、社交上以得了巨大的成功（……）。”这样的描述既简介也更为专业。建议作者对全文进行大幅度的精简，同时将语言变得更加学术化、专业化。

回应：谢谢审稿专家。已按照您的意见，将前言部分事实性的描述尽量删减，改为学术化语言，同时对全文进行精简，将正文字数控制在了 10000 字以内。

意见 2：文献综述没有提供出你的创新点，强调可以将推荐系统纳入，解决先前被试信息没被利用的问题。但是却看不出来，为什么要这样。作者在第一段有这样一句话“已有答题者数据的重要意义在于：这些过程性数据中包含了对一个特定能力答题者如何选题的经验信息，它是题目信息与答题者信息动态交互的结果，对不同答题者如何有针对性地选题具有重

要的参考价值”，在评审人看来这并不重要。作者需要告诉读者，纳入新信息是使估计更准确了？还是使题库曝光更少了？还是用时更短了？而这正是作者在后文做模拟研究在比较的内容。

回应：同意审稿专家的意见和建议，这样修改的确可以更加突出本研究的贡献和重点。已按照您的意见，增加了对已有答题者数据的重要性的阐述，加入新的角度，从数据挖掘的视角指出数据中蕴含的知识模式，合理挖掘该数据可以找到更简洁有效的选题策略，同时为改进选题策略提供指导，详见前言第一段蓝色字体修改部分。

意见 3：接上一个问题，作者需要在引言中告诉读者传统的 CAT 有怎样的不足，而加入新的推荐系统后有怎样的优势。

回应：谢谢审稿专家。已按照您的意见，在引言中强调了传统 CAT 存在的局限性，并增加了引入协同过滤推荐后可能带来的优势，不仅能够利用已有答题者数据，而且避免了复杂的计算公式和约束，同时建立了一个灵活可拓展的新框架，详见前言第一、二、五段蓝色字体修改部分。

意见 4：部分 2 选题策略中减少选题基础知识的介绍，突出你的方法的优势，并适当渗透用你的新方法后可能的优势预期。

回应：谢谢审稿专家。已按照您的意见，精简了对选题基础知识的介绍，并强调了新方法在设计时多处细节的预期优势，详见 2.2“基于协同过滤推荐的新选题策略”蓝色字体修改部分。

意见 5：研究设计虽然有所增加，但仍然偏简单，作为模拟研究 4 种模拟情境显得不够。在研究一的结果中作者说题库质量的问题在研究二中探索，为何不在研究一中探索。另外，上一次修改中，评审人提出的考虑被试分布的问题也可考虑直接在研究一中考察。总之，实际 CAT 中可增加的情境还有很多，作者可参照其它研究进一步增加。

回应：谢谢审稿专家，您提出的关于研究设计的建议和意见对未来进一步研究推荐选题策略的特点和改进方向有很大的启发性。如您所言，确实还有诸多可增加的情境，由于篇幅和文章主旨等原因未能逐一讨论。本研究的初衷是探讨将推荐系统中的协同过滤推荐用于 CAT 选题的可能性以及可行性，提出可以利用已有答题者信息的全新选题策略，经过大量预实验最终敲定模拟研究的重点就是围绕两个题库（模拟题库和真实题库），在不同测验长度（决定数据量）和传统选题策略（决定数据特点）下探讨推荐选题策略的表现，重点考察推荐选题策略在选题精度和对题目曝光率控制方面的表现。基于这样的研究重点，设计研究时不免有所取舍。以您指出的被试能力分布问题为例，当能力分布形态不同时，可能还会涉及到能力点的抽取方式的研究话题；如果谈到能力点的抽取方式则又需要考虑单批答题者数量和每人作答题目数量等交叉因素，这里的各个话题都是推荐系统或 CAT 选题中值得专门设计系列研究详加探讨的问题。鉴于目前本研究的中心，我们把这些很值得进一步研究的话题在讨论部分进行了说明，以待后续不断拓展完善，期待未来有更多的研究者开展更深入细致的研究，详见讨论倒数第二段本文局限与未来方向中的第二个方面（蓝色字体标注）。

意见 6：表 1，表 2 可直接合并。

回应：谢谢审稿专家。已将表 1 和表 2 合并。

意见 7：在研究一模拟情境丰富的基础上，研究二的重点就不在于将研究一的东西再用实证验证一遍了，而应该把研究设计的重点和书写笔墨转向使用已有真实作答这样的研究一无法做到的聚焦点上来。作者可以从已有作答信息量多少、准确性的程度来设计研究，例如已有

作答题量、作答准确程度等等。

回应：谢谢审稿专家。本文的主要目标是在 CAT 中引入推荐系统这一类数据挖掘技术，为改进选题策略提供一种全新的可能性。落实在具体研究问题上，是以如何利用他人作答数据为首个突破口，以基于协同过滤的推荐建立首个选题策略，如果在几种情境下以多种指标衡量都有不亚于传统策略的表现，得以初步验证引入推荐系统成功的可能性，本文的首要目的便达到了。下一步既可以针对 IEBR 和 DEBR 拓展考察的情境，又可以改进和提出更灵活多样的推荐选题策略，关于可深入研究的方向在讨论部分均有涉及，评审专家提出的角度都已视作最重要的研究主题，详见讨论倒数第二段本文局限与未来方向中的第二个方面（蓝色字体标注）。

第三轮

审稿人 1 意见：作者已按照审稿人的意见进行了修改，文章质量有较大提升。有几点建议可供作者参考修改，修后可发表。

意见 1：“使用协同过滤推荐完成 CAT 选题，可以避免传统选题策略复杂的计算公式和约束流程，从已有答题者数据中快速筛选出适合当前答题者作答的题目。此外，在协同过滤推荐的假设之上可以根据研究者需要加入其它规则，设计出可灵活扩展的选题策略 ，既可以侧重选题精度或题目曝光率控制，也可以在保证一定精度的情况下兼顾题库使用和测验安全。”——可以举一个例子，如何拓展，可实现什么目标。

回应：谢谢审稿专家。已在本段文字末尾加入一种设计思路和预期目标作为示例，详见引言倒数第二段蓝色字体修改部分。

意见 2：作者在前文提到：“这样设计可以扩大一次完整 CAT 对已有答题者数据的参考范围，使推荐选题策略可利用的信息更多，选题更加精准。”可见，已答题者数据的使用情况也比较关键，建议加上一个指标，用以衡量已答题者使用率的情况，反映利用了多少被试信息。被试调用率=已调用人数÷总人数。

回应：谢谢审稿专家。已根据新的建议增加了新指标“答题者调用率”，指标介绍见“3.2 评价指标”，两个模拟研究中对新指标的分析详见 3.3 和 4.2“研究结果”，已用蓝色字体标出。

审稿人 2 意见：经过两轮的修改，整体文章已经比较流畅和清晰。但评审人仍然有两点建议。

意见 1：由于前期的删减，使得当前的研究结果显得过于单薄。建议作者增加一些结果信息，例如考虑对被试人群特点，题目特点、测验特点进行分类的探讨。

回应：谢谢审稿专家。根据您的建议，再次查阅选题策略相关文献，在模拟真实情景时增加了按区分度参数大小呈现每种条件下的题目曝光率，同时按传统和推荐选题策略分类，按合并数据的轮次进行纵向比较，并与另一位专家建议增加的“答题者调用率”指标相互对照，发现了新指标与曝光率的变化规律具有一致性，并对此进行了深入探讨。同时，我们也分被试能力水平，考察了不同选题策略对不同水平被试的能力估计精度，结果发现推荐策略选题策略与传统选题策略的结果基本一致。详见“4.2 研究结果”中新增的图 1、图 2 和蓝色字体修改部分。

意见 2：作者将实证研究的部分转向了题库质量不如模拟研究这个点上，这是一个行得通的研究点。作者可以在结果中呈现，实际题库的质量是如何的，跟模拟题库的比较。但同时另外一个问题随之产生，模拟研究中同样可以模拟题库质量不佳的情况，作者为何不在模拟研究中更充分的论证。

回应：谢谢审稿专家。本文初稿中曾将模拟题库和真实题库各参数的描述性统计结果详细列出（见下表 1，最右为新加列，一审修改研究设计后使用同一原始分布参数新生成的模拟题库），之后参考第一轮审稿的专家 2 意见 4 删去此表。现在结合您的建议扩充了该部分内容，在设计部分简要说明了题库质量的差距体现在何处，详见“4.1 研究设计”蓝色字体修改部分。

表 1 各题库中题目参数的描述统计结果

	2PLM		3PLM		真实题库	当前研究一使用的模拟题库
	a 参数和 b 参数无相关	a 参数和 b 参数有相关	a 参数和 b 参数无相关	a 参数和 b 参数有相关		
$r_{a,b}$	0.001	0.409***	0.032	0.466***	0.125***	0.419***
M_a	1.207	1.196	1.221	1.191	0.948	0.974
SD_a	0.500	0.497	0.460	0.459	0.278	0.479
MAX_a	1.924	1.826	2.064	2.086	2.043	2.086
MIN_a	0.252	0.240	0.226	0.207	0.321	0.201
M_b	-0.049	-0.025	0.019	-0.040	0.599	-0.087
SD_b	0.947	0.986	0.995	0.986	0.579	0.927
MAX_b	2.804	3.223	2.982	3.152	2.084	2.825
MIN_b	-2.235	-3.313	-3.385	-3.271	-1.426	-2.925
M_c	0	0	0.257	0.265	0.247	0.185
SD_c	0	0	0.126	0.122	0.076	0.074
MAX_c	0	0	0.650	0.615	0.507	0.302
MIN_c	0	0	0.001	0.005	0.092	0.005

注：对于研究二使用的真实题库，计算 c 参数的描述统计时仅包括 3PLM 下的题目。*表示 $p<.05$，**表示 $p<.01$，***表示 $p<.001$。

关于没有在模拟研究中进一步考虑题库质量影响的问题，主要是由于本研究最想探究的是所提出的基于推荐系统的新选题策略在理论和实践两方面是否具有可行性。因此，研究主要基于推荐系统的本质特点，把考察重点放在了已有答题者数据特点（生成首批数据的传统选题策略）和数据量（测验长度）这两个最常见的影响推荐系统表现的因素上。而没有将题库质量作为一个考察因素（实际上，结合研究一和研究二的结果，题库质量也不是主要的影响因素）。为了证明这一思想的可行性，研究一尽量简化设计，在比较理想的题库质量假设下证实了这一想法理论上的可行性；研究二重在基于真实的题库，考察推荐选题策略在实践应用中多次数据迭代后的预期表现以及与传统方法的比较，结果在实践层面证实了这一方法的可行性，两个研究是从两个角度证实基于推荐系统设计选题策略这一思想的合理性和可行性。当然，作为一个新方法的探索性研究，值得研究的问题还很多，我们也同意您的建议，可以考虑更多的影响因素，通过更多的模拟研究来探讨这一方法的适用条件、应用中应该注意的问题和有待改进的地方。我们再次修改了讨论中的本文局限和未来方向，重点阐述了这部分的内容，详见讨论第 4 段蓝色字体修改部分。

第四轮

审稿人 1 意见：两点小的意见：
意见 1：数据结果的描述，文字内容太多，看上有点像讨论，有点像结论。就图表进行描述，可罗列数值，目的是提示读者注意看那些点，注意那些趋势，不需要做过多的延伸、解释和推论。
回应：谢谢审稿专家。已对结果中带有延伸性质的表述进行了精简，除了对图表的直接描述外仅保留对趋势的简要概括，详见 3.3 和 4.2“研究结果”蓝色字体标注内容；同时将一些推论和解释移至讨论，详见讨论第一段和第三段中蓝色字体修改部分。

意见 2：结果表格中呈现，或者统一中文，或者统一用方法中介绍的统计量符号。
回应：谢谢审稿专家。已将表格中指标名称全部统一为中文，详见 3.3 和 4.2“研究结果”表 1 和表 2，已用蓝色字体标出。

编委复审

编委意见：摘要字数多了点，最后两句可以考虑删除或者精简。
回应：谢谢编委专家，参考您的建议和摘要的写作要求，已删去最后两句话。