《心理学报》审稿意见与作者回应

题目：基于 CD-CAT 的多策略 RRUM 模型及其选题方法开发
作者：戴步云，张敏强，焦璨，黎光明，朱华伟，张文怡

第一轮

审稿人 1 意见：应用计算机化自适应诊断测验的形式，基于 RRUM 模型做策略诊断，这个想法很好，第一改变了策略诊断停留在纸笔测验的形式的局面；第二，没有使用研究者经常使用的 DINA 模型，而是使用了平常使用比较少的认知诊断模型 RRUM，所以文章有比较高的创新性。对于策略和知识状态的联合估计的想法也很好。
但是审稿专家认为对于这个研究，存在如下一些问题：

意见 1：有的文献引用不准确：涂冬波等人（2012）的文章作者是 4 位，为何变成 3 位？对这篇文献的概括也不全面：这篇文献的摘要强调的不仅仅基于认知诊断模型 GDD，而且还结合了拓广 Q 矩阵理论，否则无法挑选知识状态；请作者对照原文查一查；
回应：感谢审稿专家的细心指正！这是我们的疏漏，抱歉。我们已经将遗漏的作者名字补充完整。另外，第二节第二段的相应内容修改为“在 de la Torre 等人（2008）的基础上，涂冬波等（2012）基于孙佳楠、张淑梅、辛涛和包珏（2011）的广义距离判别法以及丁树良、祝玉芳、林海菁和蔡艳（2009）修正的 Q 矩阵理论……”，红色部分为修改之处。

意见 2：CD-CAT 和 CAT 一样，必须有题库。文章中的题库如何模拟（建立），才能够有比较高的判准率？文章似乎没有提及这个问题；
回应：我们仔细研究了审稿专家的意见，回答如下：MS-RRUM 对题库的 Q 矩阵并无特殊要求。因此本研究在模拟 Q 矩阵的时候，仅仅是“设每个题目在每种策略下考查每个属性的概率为 0.5。若某题在某策略下的 Q 向量的模拟结果是所有元素全为 0，则重新模拟，因为实际测验中不可能有某个题没有考查任何属性。由此先后独立地随机模拟生成两个 Q 矩阵。” “丁树良、杨淑群和汪文义（2010）指出，如果测验 Q 矩阵中包含可达矩阵 R，判准率会更高。出于这一点，以后可以设计研究来验证这个定理在多策略认知诊断中是否同样成立。若成立，则建议在编制测验时注意让 Q 矩阵包含 R 矩阵。不过从这项研究的结果看，即使没有刻意让 Q 矩阵中包含 R 矩阵，MS-RRUM 的判准率已经较为理想。”（红字内容已经补充到文章最后一节。）

意见 3：对于如何估计项目参数，作者在第三节最后一段给出一个较为简单的方法“先小规模试测，让参加试测的被试做口头报告，由此确认每个被试所使用的策略；然后用常规的 RRUM 模型的联合估计算法对每种策略下的项目参数和被试 KS 做联合估计，由此获取各题的项目参数”。这个方案的可行性还需要论证，因为小规模试测和口语报告，获得被试的解题策略，由于是“小规模”，后面使用常规的 RRUM 的联合估计，样本容量够吗？如果要达到一定规模的口语报告，似乎又很困难。如何给出解决这个问题的切实可行的方案？
回应：审稿专家的意见很中肯。要达到一定规模的口语报告，确实是很困难。为此，我们将口语报告改为做书面的策略调查。将第三节最后一段修改如下（红色部分为修改之处）：

在实际测验中，用 MS-RRUM 模型开展认知诊断，必须先建设题库，而建题库时需要

估计各试题在 MS-RRUM 模型中的各策略下的项目参数。要做这样的项目参数估计，可以开发 MCMC 程序，但较为复杂且耗时耗力。一个较为简单的方法是：先找一批被试来试测，并在测验中附一个书面调查，通过被试的回答来判断被试策略倾向；然后用常规的 RRUM 模型的联合估计算法对每种策略下的项目参数和被试 KS 做联合估计，由此获取各题的项目参数。这里需要保证，使用每种策略的被试数量都达到常规的 RRUM 模型的联合估计算法所需的最少人数。

意见 4：策略数目 M 比较大时候，ORP 要和许多 IRP 比较，CD-CAT 的诊断速度如何保证？文章模拟时仅仅针对 M=2 讨论，问题不会太大，但是考虑到实际应用，文章至少应该对此进行讨论；

回应：感谢审稿专家的意见！我们在文章最后一节增加了一段，对这个问题予以说明：

MS-RRUM 模型在理论上可以处理策略数 $M \geq 2$ 的情况，这里仅以 $M = 2$ 为例进行了模拟验证。在实际工作中，若 M 值较大，在参数估计时就需要将观察反应模式与很多种理想反应模式进行比较，计算量就大大增加，在同等计算机设备下计算所需要的时间大致与 M 值成正比。不过，随着计算机技术的迅速发展，M 值增大给多策略 CD-CAT 诊断速度带来的影响就会越来越小。

意见 5：两种策略都是包含 K 个独立属性（属性的独立结构这可以由作者介绍的知识状态的数目推断出来），这时候如果使用 K 维 0-1 向量描述知识状态，那么对应的两个知识状态的集合完全一模一样，如果每一种策略使用的先验概率基本相同（比如模拟中 w1 和 w2 不是一个为 0.7，另一个为 0.3，而是几乎相等），会不会出现策略分辨不清从而导致知识状态分辨不清的可能？

回应：对审稿专家的意见进行如下解释：

首先，根据文献综述中“在解答同一数学问题时，有的学生擅长几何学的方法，有的学生擅长代数方法，只要会其中一种方法就可以做对该数学题（陈秋梅, 张敏强, 2010）”的观点，MS-RRUM 允许各种策略下的属性不相同。策略 1（Q 矩阵 1）之下的属性是 $A_1, A_2, ..., A_{K_1}$，而策略 2（Q 矩阵 2）之下的属性是 $B_1, B_2, ..., B_{K_2}$，由此类推。 红色文字已经补充进论文 3.1 节。

其次， 在模拟验证的时候，考虑到实际工作中可能存在“多数人使用一种主流策略”的情况，这项研究以“假设所有被试中使用策略 1 的人数为 70%”为例来生成数据。实际上，对于两种策略使用的先验概率相等的情境，笔者也进行过模拟研究尝试，发现策略判准率 SMR 和认知状态判准率 PCCR 同样良好。这也从一个侧面反映了 MS-RRUM 模型的性能。为了说明问题，对这个意见所做的修改已经用红色文字补充在论文的最后一节。

意见 6：尽管文章对不同的 Q 矩阵增加下标加以区分，但是公式 14 的左边缺少标记策略的字母 m,应该补充上去；

回应：感谢审稿专家的意见！由于我们假设每个被试只使用一种策略，故公式 14 的左边的 p_{ji} 的下标里不加 m 也是可以辨识的。

意见 7：文章最后只有结论，没有讨论，建议增加一点讨论的内容。

回应：感谢审稿专家的建议！我们将最后一节的标题“6 结论”改为“6 结论、讨论与展望”，并增加了数百字的讨论内容。

审稿人 2 意见：

文章尝试解决了 CDCAT 中的多策略问题，选题具有理论价值和实用价值，使得多策略认知诊断从纸笔测验发展到了计算机化自适应测验中，具有一定创新性，但文章还存在以下问题：

意见 1：引言没有起到“引”的作用，和本文要介绍的多策略认知诊断没有呼应，转而在第二部分介绍多策略认知诊断研究综述就显得有些生硬。

回应：感谢审稿专家的指正！我们对此做如下修改：（1）将原有第二部分的第一段“目前广泛使用的认知诊断模型有一个前提假设……”挪到了引言部分，并在后面加了一段“那么，若 CD-CAT 中存在多种解题策略，如何选题才能更加高效、快速地估计出被试的解题策略倾向和认知状态，就值得探究。”

意见 2：正文 P4 最后一段作者提到涂冬波等（2012）的文章已经提出了 MSCD 方法，也使用了解题策略判准率指标，之后，也提到了刘铁川（2012）提出的 Mix-DINA 模型，而本文并未使用这两种方法进行 CDCAT 研究，而是采用了新开发的 MS-RRUM，请说明理由

回应：感谢审稿专家的意见！对于 DINA 模型，它的不足之处已经在文中有所介绍；而 Mix-DINA 模型作为 DINA 模型的衍生，并没有改变这些不足之处，所以我们没有选择 Mix-DINA 模型。涂冬波等（2012）提出的 MSCD 方法是基于判别分析，如果用它来做 CD-CAT，选题会很困难，而且常用的选题方法诸如 SHE 等很可能都不能使用，所以我们没有选择 MSCD。但审稿专家的这一意见为我们今后的研究指出了一个方向。

意见 3：作者虽提出了 MS-RRUM，但并没有对其模型性能等进行前期研究，予以检验，而是直接将该模型用于 CDCAT 测验。在不知一个模型性能如何时，使用该模型是否会存在问题？作者是如何考虑的，请予以说明。

回应：审稿专家的意见很中肯。在刚刚做 MS-RRUM 研究的时候，我们确实不知道这个模型的性能如何。所以我们在提出了该模型的参数估计方法（多策略 MAP 法）和 MSSHE 选题法之后，根据 CD-CAT 的两个指标（SMR 和 PCCR）来检验。只有当 MS-RRUM 的模型性能、多策略 MAP 法和 MSSHE 选题法的性能都良好的时候，最后才能得到良好的指标结果；只要这三者中有一个的性能不佳，就不能得到良好的指标值。也就是说，我们在一个模拟实验里同时检验了三者的性能。最后的指标值很好，这就说明了模型性能良好。为了更好地说明这个问题，我们在文章末节的第一段对此进行了解释。

意见 4：作者在 3.1 部分说道：“在多策略情境中，被试的认知状态 KS（即属性掌握模式）是跟他所使用的策略相关联的。被试能否成功解题与被试所具备的策略模式下的 KS 有关。”这里有一个很关键的问题需要厘清：如果在不同策略中所考察的属性不完全一样，例如 de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: an analysis of fraction subtraction data. Psychometrika, 73(4), 595–624 一文中表 7 给出的 Q 矩阵形式：策略 A 考察了属性 A1，A2，A3，A4，A5，策略 B 考察了属性 A3，A4，A5，A6，A7。该情况下，被试的 KS 就会在两种不同的向量组合间跳来跳去，请问此时被试的 KS 该是什么？是 α(A1，A2，A3，A4，A5)，还是 α(A3，A4，A5，A6，A7)，还是 α(A1，A2，A3，A4，A5，A6，A7)？那么，MS-RRUM 又是如何做出诊断的？请予以说明。

回应：感谢审稿专家的意见！这里需要解释的是： MS-RRUM 允许各种策略下的属性不相同。策略 1（Q 矩阵 1）之下的属性是 $A_1, A_2, ..., A_{K_1}$，而策略 2（Q 矩阵 2）之下的属性是 $B_1, B_2, ..., B_{K_2}$，由此类推。 红色文字已经补充进论文 3.1 节。 由此，只有当某被试掌握

了某种解题策略下的全部属性的时候，才可能出现您所说的问题。MS-RRUM 只能诊断出某被试在他所采用的策略下的 KS，无法诊断出他在未使用策略下的 KS。而这并不是这个模型本身的问题。根据文献综述里“在解答同一数学问题时，有的学生擅长几何学的方法，有的学生擅长代数方法，只要会其中一种方法就可以做对该数学题（陈秋梅，张敏强, 2010）”的表述，如果某学生一直采用代数策略下解答这类数学问题，任何认知诊断方法都无法估计出该学生在几何策略下的属性掌握模式（即 KS）。

意见 5：作者在进行模拟研究时，将每个策略下所考察的总属性数量规定为相等，即 $K_1 = K_2 = K$，但在模型介绍时并没有阐述是否应该有所限制，请问，MS-RRUM 能否处理每个策略下总属性数量不相等的情况呢？请予以说明

回应：感谢审稿专家的意见！MS-RRUM 可以处理每个策略下总属性数量不相等的情况。在模拟研究时设置“每个策略下所考察的总属性数量规定为相等”仅仅是为了便于举例说明。

意见 6：作者提出的多策略模型是公式（2）吗？de la Torre 和 Douglas(2008)提出的多策略模型为 $P[Y_{ij}=1 \mid \boldsymbol{\alpha}_i]=\sum_{m=1}^{M} P[\omega_{ij}=m \mid \boldsymbol{\alpha}_i](1-s_{jm})^{\eta_{ijm}} g_{jm}^{1-\eta_{ijm}}$，其中 $P[\omega_{ij}=m \mid \boldsymbol{\alpha}_i]$ 为被试选择策略 m 的概率，该模型允许被试在解答题目时，可以以一定的概率进行策略间的切换，而作者本文提出的多策略模型为 $p_{l_{mi}}=\pi_{mi}^{*}\prod_{k=1}^{K_m} r_{mik}^{*(1-\alpha_{l_{mk}})q_{mik}}$，少了 $P[\omega_{ij}=m \mid \boldsymbol{\alpha}_i]$ 表征选择某个策略的概率这一部分，请问作者是如何在 MS-RRUM 中表征被试使用某种策略的概率的？请作者对构建该模型的原理进行解释

回应：感谢审稿专家的意见！在 MS-RRUM 中，假设某个被试在整个测验中只使用一种策略，故不存在策略切换的问题。（这句话已经补充进文章末段。）而在用多策略 MAP 方法进行参数估计的时候，设置每个被试使用每种策略的先验概率都相等。

意见 7：作者在 5.2 部分第三步骤说道：“每个被试在整个测验中只使用一种策略”，这是否与多策略的本质存在矛盾？而且根据模拟结果，SMR 并非都等于 1，表明有些被试在整个测验中是使用了两种策略的，这又如何理解？若规定了每个被试在整个测验中只使用一种策略，那么为何不分成 M 个 Q 矩阵分别做单策略的认知诊断呢？请作者予以解释。

回应：对审稿专家的意见，回答如下：（1）多策略的本质是指，面对同样的测验任务，有的被试使用这种策略，有的被试使用那种策略。这既包括“整个测验中每个被试只使用一种策略”的情况，也包括“同一个被试在整个测验中使用多种策略”的情况，本研究还针对前者，这相对简单。后者更复杂，未来的研究需要考虑这种更复杂的情况，以更贴近客观实际。（红色文字已经补充进最后一段。）（2）模拟结果显示，SMR 并非都等于 1，这说明有些使用策略 1 的被试被误判为使用策略 2，而有些使用策略 2 的被试被误判为使用策略 1，这并不意味着“有些被试在整个测验中使用了两种策略”。（3）如果要“分成 M 个 Q 矩阵分别做单策略的认知诊断”，就要分两步走：先确定被试的策略，然后在该策略下估计被试的 KS。而我们开发的多策略 MAP 估计方法中，被试的策略参数和 KS 是结合起来同时算后验概率，然后选取后验概率最大的那个组合，策略和 KS 诊断一步完成。

意见 8：作者在第六部分说道：“（3）能够估计出每个被试使用每种策略作答的概率，充分

提供了策略诊断的信息，弥补了 de la Torre 等人（2008）研究的不足".审稿专家认为作者的模型和 de la Torre 等人（2008）提出的模型是不一样的，正如第 7 条意见所述，因此作者虽能得到策略诊断的信息，但还是没有解决 de la Torre 等人（2008）模型中的问题。

回应：感谢审稿专家的意见。我们已经在原文里将这句话删除。

意见 9：参考文献中有的文章并没有在正文中出现，建议予以删除或在正文中标示出来，例如：郭磊. (2014). 变长认知诊断计算机化自适应测验：终止规则、曝光控制及题库质量监控技术(博士学位论文). 北京师范大学.

回应：感谢审稿专家的细心指正！我们已经将这一条参考文献删除。

意见 10：其他意见请参加原文中批注。

回应：感谢审稿专家的建议！有一些批注之处，已经按照您的意见直接修改。还有一些批注之处，我们解释和修改如下：

批注 6：baseline parameter 这个名词，已有的国内文献均译为“难度参数”，故本文也沿用这个译法，但附上英文原词，以便于交流。

批注 8：RRUM 模型由 Hartz(2002)提出，在原文里对区分度参数 r_{ik}^{*} 是这样表述的：

> apply all the attributes when solving item i. For an examinee lacking a required attribute, her correct item response probability is proportional to r_{ik}^{*} for the attribute k she is lacking. Thus r_{ik}^{*} is the penalty for lacking attribute k, in terms of comparing the correct item response probabilities between lacking attribute k and mastering attribute k. In effect, it is a discrete, attribute-based discrimination parameter for item i.

而在涂冬波等（2012）的书中，是这样表述的：

（2）$r_{ik}^{*}=\dfrac{P(Y_{ijk}=1|\alpha_{jk}=0)}{P(Y_{ijk}=1|\alpha_{jk}=1)}$：被试缺乏属性 k 与掌握属性 k 但都答对项目的概率比，它能反应属性 k 的重要性，若其值为 0.25，则说明被试掌握属性 k 答对该题的概率是未掌握属性 k 也答对该题的概率的 4 倍，也就是说掌握属性 k 对答对该题很重要。r_{ik}^{*} 的值越小说明属性 k 越重要。它被称为项目 i 属性 k 的区分度参数，其值界于 0-1 之间；r_{ik}^{*} 越小说明项目 i 的属性 k 在正确答对项目 i 上越重要，也即该属性越能区分开答对与答错该题的被试，属性 k 有高的区分度。一个项目若有 K 个属性，则该项目有 K 个区分度参数。

所以我们想维持原文对 r_{ik}^{*} 参数的表述。

批注 9：同意您的意见。现在我们这样修改：“假设某一类任务存在 M 种策略模式”。红色字体为改动之处。

批注 12 和批注 14：Y_j 确实是向量。现在我们这样修改：“令 $P(st_j=m,\alpha_j=\alpha_{l_m}|Y_j)$ 是当被试 j 的作答向量为 Y_j 时，被试 j 的策略参数为 m 且 KS 为 α_{l_m} 的后验概率。”（红色部分

为修改之处。）

批注 13：（1）i 是题库里的编号，t 是被试 j 所做过的题的顺序。被试 j 所做过的第 t 题可能是题库里的第 i 题。故这里的表述并无冲突之处。（2）对 ω 求和确实是从 1 到 M，我们已经修改。

批注 15：在第 3 节，MS-RRUM 模型已经开发完毕。后面的模拟研究含有对该模型进行性能检测的用意。只有当 MS-RRUM 的模型性能、多策略 MAP 法和 MSSHE 选题法的性能都良好的时候，最后才能得到良好的指标结果；只要这三者有一者性能不佳，就不能得到良好的指标值。也就是说，我们在一个模拟实验里同时检验了三者的性能。最后的指标值很好，这就说明了模型性能良好。（第 6 节第一段的红字是本次修改的部分。）

批注 16：查阅了 Xu, Chang, Douglas（2003）对于 SHE 公式的表述，我们认为这里还是用大写的 Y 更合适。

批注 17：公式（8）和（11）中，m 处于连加号的下方，最后得到的 $MSSHE_i(\pi_{j,n})$ 值与 m 无关，故似乎不必将 $MSSHE_i(\pi_{j,n})$ 改成 $MSSHE_{im}(\pi_{j,n})$ 。

批注 19：这是经过前期试探以后选定的水平。从后面的模拟结果看，这两个水平选得比较合理。若短测验太短，会使得属性数较多时 PCCR 较低；若长测验太长，会导致天花板效应更明显，反而不利于探究最合适的测验长度。

批注 20：对于“每一种策略使用的先验概率基本相同”的情境，我们也用 MS-RRUM 模型做过模拟研究尝试，发现策略判准率 SMR 和认知状态判准率 PCCR 同样良好。（第 6 节倒数第二段已经补充了相应文字。）由于在客观现实中每一种策略使用的先验概率可能是不相等的，故我们在正式研究中没有设置先验概率相等。

批注 21：您所说的这种情况很复杂，这是本研究所提出的 MS-RRUM 模型所不能解决的。为此，我们决定在最后一段加这几句话：本研究假设每个被试在整个测验中只使用一种策略。但是在实际测试中，可能有的被试在不同的题目中采用了不同的策略。为此，未来的研究需要考虑这种更复杂的情况，以更贴近客观实际。

批注 23：解释同批注 15。

批注 24：和您一样，我们也猜测“选题策略应该不会受到认知诊断模型的影响”。但由于我们没有在其它模型下使用 MSSHE 选题法，故不敢轻易下结论，只好寻求较稳妥的表述。

批注 25：在选题方法的效率较高时，要达到设定的测验精度所花费的题目就较少，这样就可以将定长测验的长度设置得较短，于是节约了测题的数量。（红色文字已经补充到原文中。）

第二轮

审稿人 1 意见：作者基本上比较好地回应了审稿人提出的问题。但是还有以下几个小问题提请注意：

意见 1：从数学公式的严谨性出发，建议作者对（14）的左边做出修改，即增加下标；

回应：感谢审稿专家的指正！我们已经修改了该公式。

意见 2：如果某些被试对所有题目统统正确作答（或者统统错误作答）；这些人使用的策略无法估计，这一点应该补充说明；

回应：感谢审稿专家的指正！我们在公式（4）后面加了一段话：“值得注意的是，如果某被试答对了或答错了所有的题目，则该被试的策略无法估计。”（在论文中，这次修改的部

分均用蓝色表示。）

意见 3：在最后面一节作者表示“实际上，对于两种策略使用的先验概率相等的情境，笔者也进行过模拟研究尝试，发现策略判准率 SMR 和认知状态判准率 PCCR 同样良好。这也从一个侧面反映了 MS-RRUM 模型的性能”。审稿人认为，这是很重要的结果，建议作者补充（哪怕是作为附录补充也好）。

回应：感谢审稿专家的意见！两种策略使用的先验概率相等的情境之下的数据结果如下：

附表 1　两种策略使用的先验概率相等，*K*=4 时，不同选题方法和不同测验长度下的平均判准率

		策略判准率 SMR		模式判准率 PCCR	
方法	测验长度	平均	标准误	平均	标准误
随机	短	0.814	0.004	0.589	0.005
随机	长	0.854	0.003	0.666	0.005
MSSHE	短	0.978	0.001	0.955	0.002
MSSHE	长	0.988	0.001	0.978	0.001

附表 2　两种策略使用的先验概率相等，*K*=6 时，不同选题方法和不同测验长度下的平均判准率

		策略判准率 SMR		模式判准率 PCCR	
方法	测验长度	平均	标准误	平均	标准误
随机	短	0.841	0.004	0.477	0.007
随机	长	0.876	0.003	0.559	0.006
MSSHE	短	0.990	0.000	0.954	0.002
MSSHE	长	0.995	0.000	0.979	0.002

附表 3　两种策略使用的先验概率相等，*K*=8 时，不同选题方法和不同测验长度下的平均判准率

		策略判准率 SMR		模式判准率 PCCR	
方法	测验长度	平均	标准误	平均	标准误
随机	短	0.834	0.003	0.350	0.007
随机	长	0.868	0.003	0.412	0.007
MSSHE	短	0.988	0.001	0.908	0.004
MSSHE	长	0.992	0.001	0.935	0.004

将这 3 个表格与正文里的表 1、表 3、表 5 进行对比。总的来说，在同等属性数量和同等测验长度的条件下，当两种策略使用的先验概率改为相等之后，各方法的 SMR 和 PCCR 都略有下降，但下降幅度不大。随机法的 SMR 下降幅度在 0.070 到 0.088 之间，PCCR 下降幅度在 0.033 到 0.056 之间；SHE 法的 SMR 下降幅度在 0.004 到 0.013 之间，PCCR 下降幅度在 0.04 到 0.022 之间。可以看出，SHE 法的结果受“策略使用的先验概率”的影响很小。（这些内容已经放入附录里。）

意见 4：审稿人认为“5.2 节模拟过程”描述得太简单，一方面作者的研究问题很新，没有相关的文献支撑，读者难以掌握，细节介绍清楚可读性才会比较好；另一方面，模拟细节介绍清楚，读者才能够重复作者的试验结果，便于推广作者成果。这里至少存在一些细节应该进一步交代，比如（1）作者给出的知识状态的总数表明属性层级关系是独立型，随机模拟生成的两个 Q 矩阵是什么，能否给出？因为 CD-CAT 的题库建设是保证认知诊断成功的物质

基础，不交代似乎无法进行其他的流程；（2）被试的知识状态如何生成？其分布如何？不交代无法说清楚模式判准率；上面两个问题是相互联系的：比如知识状态中有的仅仅只有一个 1，而模拟生成的 Q 矩阵中没有这一类仅仅包含一个属性的题目，于是这样的被试的知识状态很难判准；如果这种被试所占比重很大，那么知识状态的判准率一定很低。按照这个逻辑，作者所说的“MS-RRUM 模型对 Q 矩阵并无特殊要求”就值得商榷。

回应：感谢审稿专家的意见！（1）由于我们设置了属性数量为三个水平（4、6、8），而且在“每种条件下，独立重复模拟 20 遍”，每次模拟时要生成 2 个 Q 矩阵，故在做研究时实际模拟生成了大量的 Q 矩阵。因此，不便一一给出。在属性数量为 4、6、8 时，我们分别选取了某次模拟时生成的两个 Q 矩阵，放在一个单独的 EXCEL 文件里（提交修改稿时作为附件上传），请审稿专家审阅。

文中对题库规模和 Q 矩阵模拟的部分，现在这样修改（其中，红色部分是上次改的，蓝色部分是这次新增的）：

“根据 Stocking（1994）的阐述，题库的大小应该在测验长度的 12 倍以上。由此，在这项研究中，题库的规模 *I* 就设定为长测验长度的 12 倍，具体为：*K*=4 时，240 题；*K*=6 时，360 题；*K*=8 时，480 题。每次模拟时，被试数量为 1000 人。”

“设每个策略所考查的总属性数量相等，即 $K_1 = K_2 = K$，则有 $L_1 = L_2 = 2^K$，记为 L。设每个题目在每种策略下考查每个属性的概率为 0.5，由此 Q 矩阵服从均匀分布。具体做法是：在模拟第 *i* 题是否考查属性 *k* 时，由系统随机生成一个（0,1）之间的数字。如果这个数字大于等于 0.5，就认为第 *i* 题考查了属性 *k*，将 Q_{ik} 设置为 1；否则，就认为第 *i* 题没有考查属性 *k*，将 Q_{ik} 设置为 0。若某题在某策略下的 Q 向量的模拟结果是所有元素全为 0，则重新模拟该题，因为实际测验中不可能有某个题没有考查任何属性。由此先后独立地随机模拟生成两个 Q 矩阵。”

（2）被试的认知状态的模拟为：“假设所有被试中使用策略 1 的人数为 70%，每个被试在整个测验中只使用一种策略；在每种策略之下，每种 KS 的先验概率均相同，即被试在特定策略下的掌握矩阵服从均匀分布。被试的 KS 矩阵的模拟方法与 Q 矩阵类似，只是不需要最后一步，因为客观存在不掌握任何属性的被试。”（蓝色部分是这次新增的。）

当属性数 K=4 时，每道题可能考查的测验 q 向量共有 $2^4 - 1 = 15$ 种情况；此时的题库规模为 240 题，平均每种 q 向量有 $240/15 = 16$ 道题。当属性数 K=6 时，每道题可能考查的测验 q 向量共有 $2^6 - 1 = 63$ 种情况；此时的题库规模为 360 题，平均每种 q 向量有 $360/63 \approx 5.7$ 道题。当属性数 K=8 时，每道题可能考查的测验 q 向量共有 $2^8 - 1 = 255$ 种情况；此时的题库规模为 480 题，平均每种 q 向量有 $480/255 \approx 1.9$ 道题。由于被试的 KS 与 Q 矩阵都是服从均匀分布，故可以认为每种 KS 都有与之对应的题目来考查，由此保证了 KS 的判准率 PCCR 不会太低。而模拟的结果也表明 PCCR 确实不低。

审稿人 2 意见：作者对审稿人提出的修改意见进行了较为详尽地回答，并在修改稿中相应位置做了修改，使文章质量有了较大提升，基本达到发表要求。但还有一个修改时出现的问题需要厘清： MS-RRUM 为补偿模型，丁树良，汪文义，罗芬(2012). 认知诊断中 Q 矩阵和 Q 矩阵理论一文中的定理 1 指出：若采用 0-1 评分方式且属性对认知任务所起的作用是非补偿

连接的，则期望反应模式集合与知识状态集合建立起双射(bijective)的充分必要条件是可达阵 R 是 Qt 的子矩阵。因此，是否有必要在 RRUM 中考察 R 阵的作用需要考虑清楚。

回应：感谢审稿专家的指正！经过查阅文献，发现丁树良等人的文章中的结论的前提是属性之间非补偿。对于补偿型，优良测验蓝图的设计的原理还要仔细讨论。因此，在本文的倒数第四段，我们做了相应修改（蓝色部分为本次修改之处，红色部分为上次修改之处）：

MS-RRUM 模型对 Q 矩阵并无特殊要求。虽然有学者（丁树良，杨淑群，汪文义, 2010; 丁树良，汪文义，罗芬, 2012）指出，对于属性之间没有补偿关系的模型（如 DINA），如果测验 Q 矩阵中包含可达矩阵 R，判准率会更高。但对于允许属性之间有一定补偿作用的模型，优良测验蓝图的设计的原理还要仔细讨论。而 RRUM 正是一个补偿模型（Feng 等, 2014）。基于这一点，以后可以设计研究来验证这个定理在 MS-RRUM 模型中是否同样成立。若成立，则建议在编制测验时注意让 Q 矩阵包含 R 矩阵。不过从这项研究的结果看，即使没有刻意让 Q 矩阵中包含 R 矩阵，MS-RRUM 的判准率已经较为理想。

第三轮

审稿人 1 意见：作者经过两次修改以后，审稿人认为比较好地回应了审稿人的意见。只是修改稿 5.2 节有两处表述似乎不太规范，即“由此 Q 矩阵服从均匀分布”和“即被试在特定策略下的掌握矩阵服从均匀分布”。是否将 Q 矩阵服从均匀分布改为 Q 矩阵的元素服从 p=0.5 的二项分布比较合适？仅供参考。当然作者可以使用自己认为更加准确的表达。审稿人认为已经达到发表的水平。

回应：感谢审稿专家的意见！经过查阅文献以及反复思考和讨论，作者们认为“设每个题目在每种策略下考查每个属性的概率为 0.5”和“每个被试在整个测验中只使用一种策略；在每种策略之下，每种 KS 的先验概率均相同”两句话已经分别把测验 Q 矩阵和被试掌握矩阵交代清楚了。如果再增加文字，反而容易引起争议。所以作者们把“由此 Q 矩阵服从均匀分布”和“即被试在特定策略下的掌握矩阵服从均匀分布”都删除了。

（本轮没有审稿人 2 的意见。）

主编终审

意见：Can we combined Tables 1 to 6 into a larger table with all the results in one place for easy comparison?

回应：Thank you for your opinion. We think maybe combining the six tables into two tables is more proper, thus we do it.