《心理学报》审稿意见与作者回应

题目：认知诊断 CAT 中具有非统计约束选题方法的比较
作者：毛秀珍，辛涛

第一轮

审稿人1意见：认知诊断的计算机化自适应测验是当前心理测量的研究前沿问题之一。研究者抓住最新的研究问题，阅读并运用最新的研究资料对认知诊断ＣＡＴ中具有非统计约束（主要是内容约束和项目曝光率）的选题方法进行了模拟比较，对认知诊断ＣＡＴ选题策略的发展和实际应用均具有很好的指示作用。通读全文，本人认为还存在以下不足：

意见 1：引言部分请作者将 CD-CAT 与 CAT 的概念进行详细说明。为何要进行 CD-CAT，作者阐述并不到位，CD-CAT 的出现并不仅仅是因为有了 CD，自然就有 CA-CAT。以及为何要讨论非约束条件下的选题策略等请作者结合测验的实践进行讨论。

回应：根据专家的意见，已在引言第一、二段补充了“CAT 与 CD-CAT 的概念”以及“研究 CD-CAT 的原因”；在第三段结合实践补充了“CD-CAT 中讨论非统计约束的意义”。具体请参见修改稿中红色字体部分。

意见 2：认知诊断的模型类型繁多。研究选择 Fusion 的融合模型来模拟被试作答，虽然作者提及该模型更符合实际情况，但另一方面融合模型的估计参数较多，跟其它简单的认知诊断模型相比较，它并不是一个很简洁的模型。请作者从模型简洁的角度来说明选择这个模型的原因。

回应：首先我要纠正初稿中关于模型名称的错误。融合模型（The Fusion Model）的表达式如下：

$$P(Y_{ij}=1|\alpha_i,\pi_{jk},r_{jk})=\prod_{k=1}^{K}[(1-\pi_{jk})^{\alpha_{ik}}\cdot r_{jk}^{1-\alpha_{ik}}]^{q_{jk}} \quad (1)$$

统和模型（The Unified Model）（Stout & Roussos, 1995）的表达式如下：

$$P(Y_{ij}=1|\alpha_i,\theta_i,\pi_{jk},r_{jk},c_j)=\prod_{k=1}^{K}[(1-\pi_{jk})^{\alpha_{ik}}\cdot r_{jk}^{1-\alpha_{ik}}]^{q_{jk}}\cdot P_{c_j}(\theta_i) \quad (2)$$

其中 π_{jk} 和 r_{jk} 分别表示被试掌握和未掌握属性 k 时在项目 j 上正确运用该属性的概率，即 $\pi_{jk}=P(Y_{j\bullet k}=1|\alpha_{\bullet k}=1), r_{jk}=P(Y_{j\bullet k}=1|\alpha_{\bullet k}=0)$，其值不随被试的不同改变。

统和模型的参数不可识别。因此，Hartz（2002）将其参数进行改写，得到如下重参化统和模型（The Reparameterized Unified Model，RUM）

$$P(Y_{ij}=1|\alpha_i,\pi_j^*,r_{jk}^*,c_j,\theta_i)=\pi_j^*\prod_{k=1}^{K}r_{jk}^{*(1-\alpha_{ik})q_{jk}}\cdot P_{c_j}(\theta_i) \qquad (3)$$

另外，Templin, Henson, & Templin(2008)还将 $P_{c_j}(\theta_i)=1$ 时的 RUM 称为“The reduced Reparameterized Unified Model”。

本文采用重参化的统和模型的原因有以下几点。第一，该模型比常用的 DINA 模型更一般，而且有研究表明在处理实际数据情况时它比 DINA 模型更拟合数据（Rom a' n, 2009; Jang, 2005）。第二，Hsu, Wang & Chang（2013）在同等条件下运用 DINA 和 RUM 开展模拟研究的结果发现，RUM 的测量精度没有 DINA 模型的高，平均测验长度也更长。另外，基于一些采用 DINA 和 RUM 的研究论文如“认知诊断计算机化自适应测验的项目增补—以 DINA 模型为例”，“认知诊断计算机化自适应测验选题策略的研究”以及“在较短认知诊断计算机自适应测验中平衡属性测量程度”，我们推测简单的 DINA 模型在模拟研究中结果会比更复杂的 RUM 的结果更好。因而，本文欲了解更差的结果会是什么样的，于是选用了该模型。

意见 3：延续上述问题在第”5 结果”部分提到“根据融合模型的特点，项目考查的每个属性都将影响正确作答概率，掌握属性个数较少的被试比掌握属性个数更多的被试在大部分项目上的作答表现出的不确定性更高，从而使得他们的判准率更低。”作者提出出现这一现象的原因是由融合模型的特点所造成的，那么采用其它的模型是否不会出现。其次，请从其它的角度，例如属性结构的特点和研究设计等方面进行解释。

回应：我们在解释这一结果时思考不够深入。根据专家的建议，已在修改稿中进行修改和补充，请参见4.2.2中红色字体部分。

意见 4：文中作者提出，如“在 MC-RT 和 MC-RPG 方法中为使测量精度不显著受到项目曝光控制的影响，令 等于 2。最后，MC-PP 方法中 设置为 0.8，h 等于 10。”这些固定值只与其中部分方法有关，那么固定值的设置是否会影响其结果，请解释。

回应：首先，β 值越大，MC-RT 和 MC-RPG 方法的测量精度越高，项目曝光越不均匀（这在介绍 RT 方法时有说明）。其次，MC-PP 方法中将知识状态最大后验概率 $pp_{\max}$（Hsu et al., 2013）视作被试测量精度的标准，将 $pp_{\max}$ 设为 0.8 意味着当测量精度达到 0.8 时才控制项目曝光均匀性。本文采用这两个标准的原因是：Monte Carlo 方法本身可以在一定程度上提

高项目曝光均匀性，因而对其他方法在曝光控制方面的要求有所降低。由于β和$pp_{\max}$起着平衡测量精度和项目曝光均匀性的作用。于是，当更关心测量精度时，β和$pp_{\max}$可设置大一点；反之亦然。最后，MC-PP 方法中 h 等于 10，指从曝光率最小的 10 个项目中随机选择。作者尝试过不同的 h 值，发现 h 的值对结果几乎没有影响。

意见 5：文中表 2 的“0”表示一个属性都没有考察，即属性完全没有掌握。它的现实意义很小，作者为何要讨论这种模式？而且，一个属性均没有考察的被试量在 S1 和 S2 中均超总人数的 10%，请予以解释。

回应：“0”表示一个属性都没有掌握，这样的被试如同掌握所有属性的被试一样在总体中的比例较小。但在知识状态先验分布未知的情况下通常假设其服从均匀分布。因而，本文没有剔除掉这类被试。或许，后续研究可以假设掌握属性个数服从正态分布的条件下模拟被试。另外，S1 和 S2 属性结构下分别有 7 类和 8 类知识状态，所以一个属性都没有掌握的被试会占到 10%以上。修改稿在讨论部分对这点进行了说明。

意见 6：研究主要是讨论非统计约束条件下的选题策略，而且主要是两个非约束条件：内容约束和项目曝光率。首先这两者之间有何关联，例如内容平衡了，项目曝光率是否也会获得平衡，或是项目曝光率均匀但是内容控制又失衡了等。其次，作者能否将研究结果（在内容约束和项目曝光率控制两个条件下）和只控制其中一个条件的结果进行适当的比较。

回应：“内容平衡”是指测验满足一些内容约束，以保证测验信、效度；“曝光控制”为提高项目调用均匀性，以保证测验安全。一般而言，满足内容约束会稍微降低测量精度，项目曝光控制也会降低测量精度。二者是独立设置的约束条件，不会相互影响。即“内容平衡了，项目曝光也会获得平衡”，或是“项目曝光率均匀但是内容控制又失衡了”等。根据专家的建议，我们增加了在 S3 和 S4 两种属性结构下仅满足内容平衡、仅曝光控制的实验结果，并将它们与同时具有两类约束的结果进行适当比较。选用这两类结构的原因是：S1 和 S2 中同时具有内容平衡和曝光控制条件下个各方法的判准率有明显的差异，如果在 S1 或 S2 下实验可能造成仅内容约束、仅曝光控制的测量精度时各方法的测量精度没有明显差异；其次，S4 中有 4 个属性是独立的，意味着 S4 和 S5 结构类似。

意见 7：研究模拟了五种属性结构（每种结构均含 6 个属性），在每种属性结构下生成 480 个项目，构成一个题库。也就是，说每个题库只包含一种属性结构，在实际测验当中，一个题库不可能只包含一种属性结构，可能会包含多种属性结构。作者为何要进行这种研究设计，请予以解释。

回应：在实际测验当中，一个题库不可能只包含一种属性结构，但包含多种属性结构的题库往往包含更多属性。本文只讨论了 6 个属性，Leigthon 等（2004）基于 6 个属性提出了 4 种结构，因而本文在此基础上构建单一属性结构的题库。另外，本文采用这样的设计还考虑到：第一，几乎所有关于认知诊断 CAT 选题策略的研究都采用独立属性结构进行模拟研究（Cheng, 2009, 2010; Chen, Xin, Wang & Chang, 2012; Hsu, Wang & Chen, 2013; Mao & Xin, 2013, 等等）。第二，讨论属性层级结构下的模拟实验几乎都在各种属性结构下分别进行研究（涂冬波，蔡艳，戴海琦，2013；孙佳楠，张淑梅，辛涛，包钰，2011；吴智辉，2008；尚智勇，2008）。第三，在不同属性结构下运用相同方法选题可以比较不同属性结构对选题结果的影响。修改稿将在讨论部分对此进行讨论。

意见 8：认知诊断是教育、心理测量学的研究前沿，而认知诊断的 CAT 更是一个新的研究问题与实践。虽然与以往研究设计不同，即从内容约束和项目曝光率两个控制条件来讨论选题策略，但最终这些选题策略均来自于项目反应理论下的 CAT。例如，在具体选择项目时

依然依据最大信息量（MI）等。这样一来，比如只是考虑项目的曝光，那么项目的属性的曝光情况又如何呢。可是，我们现在是讨论认知诊断，是从项目所要测量的属性入手来讨论被试的属性掌握模式。脱离现有 IRT 的 CAT 选题策略，根本性地从认知诊断的角度来讨论 CD-CAT 的问题，我想这才是最基本的问题，请作者考量。

回应：专家的意见很有深度。首先，项目选择指标虽然是从 IRT 理论下引入，但是认知诊断模型下的信息量与 IRT 的信息量的计算方法完全不同。本文选用的是认知诊断模型下 Kullback-Leibler 信息量及其在考虑知识状态先验分布情况下的后验加权 KL 信息量选题。其次，测验是以项目的形式出现，影响题库利用率和测验安全的重要因素是项目出现的次数过多。虽然如同专家指出属性曝光过多也会影响测验安全。但是，我们认为认知诊断测验中准确测量被试是否掌握每个属性应该更为重要，而且属性往往以不同形式出现。因而，考虑项目曝光控制可能比属性曝光控制更有意义。另外，本文要求考察每个属性的项目数不少于 4 个能保证对每个属性测量准确。今后实验或实践只需要加上考察每个属性的项目个数的上限就可以在一定程度上降低属性的曝光情况。CD-CAT 选题策略的研究一方面可以再传统 CAT 相关研究基础上进行推广，另一方面从 CD-CAT 本身特点出发探讨选题策略。文稿在讨论部分对此有所说明。

……

审稿人2意见：请尽量将在论文中增加第2.3点中的内容。另，本人对研究本身尚有一些疑问：

意见1：作者在文中提到“融合模型不仅能区分不同属性掌握模式的被试正确作答同一项目的概率，又能区分同一被试在包含相同属性的不同项目上的作答”，这一特性只有当模型中包含有 P 部分时才成立，而当假设 P=1时，此特性就不存在了。在此种假设条件下，该模型是否比 DINA 等其它模型更有优势呢？

回应：谢谢专家的意见。首先，我们调整了文章内容，使文3.3部分表述更详细、更清楚。其次，我要纠正初稿中关于模型名称的错误。融合模型（The Fusion Model）的表达式如下：

$$P(Y_{ij}=1|\alpha_i,\pi_{jk},r_{jk})=\prod_{k=1}^{K}[(1-\pi_{jk})^{\alpha_{ik}}\cdot r_{jk}^{1-\alpha_{ik}}]^{q_{jk}} \qquad (1)$$

统和模型（The Unified Model）（Stout & Roussos, 1995）的表达式如下：

$$P(Y_{ij}=1|\alpha_i,\theta_i,\pi_{jk},r_{jk},c_j)=\prod_{k=1}^{K}[(1-\pi_{jk})^{\alpha_{ik}}\cdot r_{jk}^{1-\alpha_{ik}}]^{q_{jk}}\cdot P_{c_j}(\theta_i) \qquad (2)$$

其中 π_{jk} 和 r_{jk} 分别表示被试掌握和未掌握属性 k 时在项目 j 上正确运用该属性的概率，

即$\pi_{jk}=P(Y_{j\bullet k}=1|\alpha_{\bullet k}=1), r_{jk}=P(Y_{j\bullet k}=1|\alpha_{\bullet k}=0)$，其值不随被试的不同改变。

统和模型的参数不可识别。因此，Hartz（2002）将其参数进行改写，得到如下重参化统和模型（The Reparameterized Unified Model，RUM）

$$P(Y_{ij}=1|\alpha_i,\pi_j^*,r_{jk}^*,c_j,\theta_i)=\pi_j^*\prod_{k=1}^{K}r_{jk}^{*(1-\alpha_{ik})q_{jk}}\cdot P_{c_j}(\theta_i) \qquad (3)$$

另外，Templin, Henson, & Templin(2008)还将$P_{c_j}(\theta_i)=1$时的 RUM 称为“The reduced Reparameterized Unified Model”。

针对专家提到的“融合模型不仅能区分不同属性掌握模式的被试正确作答同一项目的概率，又能区分同一被试在包含相同属性的不同项目上的作答”，这一特性只有当模型中包含有 P 部分时才成立，而当假设 P=1时，此特性就不存在了。”，我们认为当假设 P=1时，此特性不是不存在了，只是没有 P 不等于1时那么显著。原因分析如下：

假设测验考察4个属性，项目 j 考察属性1和3。当 P 部分等于1时，RUM 将所有被试在该项目上的反应分为4类。即除属性2和4外（被试是否掌握属性2和4不会影响作答概率），（1）若被试只掌握属性1，则其正确作答的概率为$\pi_j^*\cdot r_{j3}^*$；（2）若被试只掌握属性3，则其正确作答的概率为$\pi_j^*\cdot r_{j1}^*$；（3）若被试掌握了属性1和3，则其正确作答的概率为π_j^*；（4）若被试没有掌握属性1和3，则其正确作答的概率为$\pi_j^*\cdot r_{j1}^*\cdot r_{j3}^*$。当 P 部分不等于1时，每个被试成为1类。同样，对该项目，DINA 模型却把被试分为2类，即不考虑属性2和4的掌握情况，掌握了属性1和3的被试成为1类，其余的成为1类。而且，项目考察的属性越多，RUM 越能将多类不同知识状态的被试区分开来。于是，当假设 P=1时，RUM 的特性不是不存在了，只是没有 P 不等于1时那么显著而已。

另外，对考察相同属性的项目 a 和 b。因为项目不同，项目参数也随之不同，RUM 中同一被试正确作答项目 a 和 b 的概率不同。而 NIDA 模型的参数决定于属性，与项目无关，因此，相同被试在考察相同属性的项目 a 和 b 上正确作答的概率相同。由此可见，与 DINA 和 NIDA 模型相比，RUM 更符合实际。

但是原文的表达不准确，我们已修改为“RUM 中同一被试正确作答包含相同属性的不同项目的概率 P 不同。另外，项目考察的属性越多，它越能区分更多不同 KS 的被试正确作答该项目的概率 P。”

意见 2：在 CAT 中项目的曝光率和利用率是非常重要的两个问题，而你的研究结果中虽然有提到这两个指标，但并没有做具体的报告。如果能报告过度曝光的项目比例，利用率过低

的项目比例及题库的利用率等信息可能更能说明问题。
回应：表 4 中 NU 代表没有使用的项目的个数，它反映了题库利用率；NO 表示曝光率大于 0.2 的项目个数，反映了过度曝光项目的个数。这两个指标没有专家提出的题库利用率和过度曝光项目的比例那么清楚。因此，修改稿将原来的两个指标分别修改为“UR：题库利用率”和“NOR：曝光率大于 0.2 的项目比例”。另外，从曝光率的四分位数可以了解“利用率过低的项目比例”的情况。因为项目的期望曝光率为 25/480=0.052，那么曝光率过低的标准怎么定义，是 0.01 还是 0.02，还是有其它标准？目前未曾有研究报告“利用率过低”的标准，因此，修改稿没有加上“利用率过低项目的比例”。

意见 3：任何一种 CAT 方法都不能同时达到高精度、高效率及低曝光率的目的，作者在本文中指出“综合各项指标，MC-MPI 方法的表现最好，MC-RT 方法次之。”这一结论作者是如何权衡的，是否有具体的考量或计算公式呢？
回应：作者在作时确实没有采用一个统一的指标进行比较，只是在测量精度和曝光控制都表现较好的 MC-RT 和 MC-MPI 基础上对比得到结论。根据专家建议，我们通过查阅文献，采用了陈平，丁树良，林海菁和周婕（2006）以及刘珍，丁树良，和林海菁（2008）中介绍的统一量纲的加权求和指标（修改稿 4.1.6 中红色字体部分），并将结果补充到表 5 和表 7，同时将原来的结论修正为采用统一量纲比较后的结论。

……

第二轮

审稿人1意见：《认知诊断 CAT 中具有非统计约束选题方法的比较》一文已经针对第一轮审稿专家的意见进行了认真的修改，并逐项进行了说明。此外，还补充比较了只在一个约束条件下（内容满足或曝光控制）的选题策略情况。文稿较前稿具有了很好的阅读性和流畅性。但还存在以下几点疑惑，请作者再进一步加以修改。

意见 1：“第 4.1.4 知识状态的估计方法”指出采用 MLE 法估计考生（为与全文描述一致，请修改成“被试”）的知识状态”。这里 MLE 是用于估计被试的能力，还是知识状态。请研究者进行清晰阐述。

回应： 答：请参见[批注 7] 的回答。
意见 2：其它需要修改之处，详见附件。

回应：下面是对文中标注的具体回答和修改方案。
1.[批注1]:已补充对应的参考文献。

2. [批注2]:在这个例子当中，没有计算机自适应化的认知诊断测验也能做到这一点。关于 CD-CAT 的必要性，请作者列举相应的参考文献；关于它的实用性，请列举在教育或心理或社会评价等领域的相关实例。

答：CD-CAT 诊断结果的准确性与属性的定义息息相关。虽然授课教师也可以根据学生的表现大致推知考生是否掌握某些知识点，但这个推断与认知诊断存在根本差异。认知诊断中属

性包括的内容通常由专家讨论决定。它运用基于概率模型的方法进行推断，其结论的准确性较一般观察诊断更高。这个例子表述不够恰当和完整，已根据 Tatsuoka 和 Tatsuoka（1997）的研究做了相应修改。根据专家意见将原文修改为“例如，Tatsuoka 和 Tatsuoka（1997）采用 CD-CAT 施测分数加法测验证明它在测验结束能即时且较准确的判断考生是否掌握测验考察的每个属性（例如“从分数中分离出整数部分”、“通分”、“化简”），估计考生的知识结构，并能有效地指导补救教学。因此，CD-CAT 在教育测量领域得到越来越多的关注（Cheng, 2009; McGlohen & Chang, 2008; Xu, Chang, & Douglas, 2003; 林海菁, 丁树良, 2007; 涂冬波, 2009, 等等）。”。另外，目前很难找到包括大量项目且属性已经标定的题库。目前，除 Tatsuoka 和 Tatsuoka（1997）外几乎没有 CD-CAT 在教育或心理或社会评价等领域的相关实例。但是有大量仅认知诊断方法的研究实例或 CAT 的研究实例。因此，几乎所有 CD-CAT 都采用模拟研究，开展实证研究将是今后的发展趋势。

3.[批注3、4和5]:标题过于简洁致使表述不够清楚，已根据专家意见做了修改。

4.[批注6]：每次模拟实验计算机运行多少次？例如 S1题库，MC-IE 选题策略，同一批被试只模拟一次，获得各评价指标；还是模拟若干次，再计算各指标的均值等。

答：每种条件下只模拟一次实验。例如 S1题库，MC-IE 选题策略，同一批被试只模拟一次，获得各个评价指标的值。我们曾经在同种实验条件下做过多次模拟实验，发现多次实验结果几乎没有差异。另外，大部分 CD-CAT 研究（Cheng, 2009, 2010; Mao & Xin, 2013; Hsu, Wang, & Chen, 2013; Wang, Chang, & Hunber, 2011等）未见报道采用“同种实验条件下开展多次重复实验，然后取各指标的均值”。因而本文采取在每种实验条件下开展一次模拟实验获得各指标的值。已在文中补充说明。

5.[批注7]：MLE 是用于估计被试的能力值还是知识状态？请作者具体说明。

答：MLE 方法用于估计被试知识状态。当假设各类知识状态服从均匀分布时，MLE 与 EAP 估计方法是一致的。已在文中进行相应修改。

6.[批注8和9]:关于文中的一些笔误，已做修改。谢谢专家的细心审阅，我今后在写作时一定会更加细致和用心。

审稿人 2 意见：有一个新的问题，作者在第 4.2.1 节第二段有一个解释‘造成这样的结果可能与研究设计、属性结构以及模型等多种因素有关’，如果真如作者所说是与实验设计有关，则本文的研究价值和意义就存在很大的局限了，其结果是否具有科学性值得怀疑。建议作者仔细斟酌！
意见 1：批注 1：如果这种现象的出现真如作者所说是与实验设计有关，则本文的研究价值和意义就存在很大的局限了，其结果是否具有科学性值得怀疑。

回应：关于文稿“4.2.1”中探索“S4 和 S5 结构下属性掌握模式判准率下降的原因“时发现“在无结构和独立结构下掌握属性个数较少的那些 KS 的判准率较低，掌握属性个数较多的那些 KS 的判准率更高。”我们在初稿写作时仅从模型特点解释这些结果，当时评审专家一指出那

样解释很牵强，并建议从其它角度进行解释，如属性的结构和整体的研究设计等。由于很难准确把握出现这些结果的真实原因，我们根据专家的意见做了修改。修改稿中“研究设计”本意指的是其它实验条件（比如被试知识状态服从其它分布），并非指研究设计不恰当或者实验条件不具有实际意义。因为，当知识状态分布未知时假设其服从均匀分布是 CD-CAT 研究中惯用的方法（Cheng, 2009, 2010; Mao & Xin, 2013; Hsu, Wang, & Chen, 2013, 等等）。当然，今后有待在其它实验条件下（如假设知识状态从服从多变量正态分布的变量中根据临界值分割为二分变量获得）进行研究。这里表述不恰当，经过思考后将原文修改为“造成这些结果可能与 API（指项目属性模式）和 KS（指知识状态）的分布、属性结构以及模型等多种因素有关”。

……

第三轮

审稿人 1 意见：

意见 1：在 4.1.3 部分，作者提到“每种实验条件下模拟一次实验，因此在五种属性结构下分别运用五种方法选题，一共进行了 25 次模拟实验”。为何只模拟一次，请给予解释。

回应：我们曾在退修二中对采取“每种实验条件下模拟一次”的原因进行解释。首先，我们曾经在相同实验条件下做过多次模拟实验，发现多次实验结果几乎没有差异。其次，大部分 CD-CAT 研究（Cheng, 2009, 2010; Mao & Xin, 2013; Hsu, Wang, & Chen, 2013; Wang, Chang, & Huebner, 2011 等）未见报道采用“同种实验条件下开展多次重复实验，然后取各指标的均值”，即它们均在相同实验条件下开展一次模拟实验获得各个评价指标的值。鉴于上述原因，本文采取“在相同实验条件下模拟一次”的实验设计合理，不会影响研究结果和结论。

意见 2：虽然作者对研究结果做了清晰的表述，但在讨论部分对于出现这种结果的原因，并没有进行过多深入的分析与讨论。比如“相同方法在不同属性结构下项目曝光率分布一致，即各种方法项目曝光均匀性不受属性结构的影响…..”为什么会这样，请作者在讨论部分给予分析。

回应：根据专家的建议，已在文稿讨论部分对此进行说明和讨论。具体而言，在结论与讨论部分第一段第十二行增加如下内容“另外，相同方法在不同属性结构下项目曝光率分布一致，不同方法在相同属性结构下的表现具有明显差异。这表明，这表明，相同方法在不同条件下具有稳定的曝光率分布，因而容易推广到其它实验条件；不同方法项目曝光率分布的差异源自方法本身的差异，不受实验条件的影响。”。

……