《心理学报》审稿意见与作者回应

题目：不同认知结构被试的测验设计模式
作者：彭亚风; 罗照盛; 李喻骏; 高椿雷;

第一轮

审稿人 1 意见：

意见 1：本文探讨了针对不同认知结构的被试合适的测验设计，将 IRT 测验中的相应方法进行了改进，具有较强的创新性。但是，由于测试前被试认知结构及其分布都是未知的，本研究对于测验编制和题库构建的借鉴意义并不明显。希望作者能够认真论证本研究的实际意义。

回应：首先，感谢审稿专家对本文研究方法的肯定。其次，我们将从以下两个方面阐述本研究的意义：第一方面，从理论上说，本研究旨在通过为不同认知结构的被试个体探索其最佳的测验设计模式，从而达到为被试群体优化诊断测验设计模式的目的。在认知诊断评价中，首先需要确定测量目标。一旦目标确定，则该目标领域内的属性及其层级关系便也随之确定，所有可能的认知结构类型便也固定了。此时不论被试群体的分布形态如何，在不改变目标领域属性的前提下该群体的认知结构类型是固定不变的。因此，探索不同认知结构的最佳测验设计模式，可以明确认知结构与项目之间的关系，从而针对被试群体设计适合每个个体的最佳测验设计模式，进而实现对群体施测时的测验优化。第二方面，从实践上说，根据群体最优的测验设计方案可以在目标测验领域内构建有针对性的题库，以优化对被试群体的诊断，避免了盲目追求建设大容量的题库，节约题库建设成本。

意见 2：考前被试的知识状态/认知结构是未知的。即使我们知道不同知识状态的被试应该做不同的测验，我们也没有办法为他们选择这样的测验。同理，考前被试的知识状态分布也是未知的，按照均匀分布构造出来的题库在真实测验中未必能够保持较好的特性。

回应：本文的主要目的是通过对不同认知结构类型个体的最佳测验模式的探索，以达到群体诊断最优。因此，为考虑被试分布的完备性，保证在目标领域范围内的每种认知结构都有，且每种认知结构的人数足够多，我们采用了均匀分布构造题库。通过逻辑分析，我们认为被试分布形态对本研究的结论应该不会产生实质影响。当然，为了谨慎起见，我们根据专家意见，模拟了不同被试分布状态下的题库使用情况，如下表所示。

定长情况下 3 种被试分布时 3 种题库的使用情况

分布形态	题库	χ^2	$\hat{T}$	*Max. er*	*Min. er*	$er \geq 20\%$	never used%	PMR
正态分布	题库 1	49.640	0.458	0.966	0	31.467	9.35%	0.953
	题库 2	63.582	0.549	0.990	0	30.567	28.42%	0.969
	题库 3	150.090	0.472	0.960	0	30.400	52.83%	0.978
正偏态分布	题库 1	52.508	0.477	0.986	0	30.767	10.84%	0.930
	题库 2	66.076	0.566	0.994	0	30.433	25.95%	0.961
	题库 3	158.560	0.496	0.986	0	29.767	50.15%	0.971
负偏态分布	题库 1	51.570	0.470	0.925	0	31.800	10.78%	0.969
	题库 2	63.553	0.549	0.964	0	32.733	33.35%	0.976
	题库 3	162.004	0.505	0.925	0	31.967	56.30%	0.986

不定长情况下 3 种被试分布时 3 种题库的使用情况

分布形态	题库	平均测验长度	*Max.*	*Min.*	$er \geq 20\%$	never used%	PMR
正态分布	题库 1	15.385	0.955	0	21.533	13.52%	0.955
	题库 2	14.180	0.982	0	21.233	23.31%	0.962
	题库 3	12.983	0.958	0	17.733	58.45%	0.967
正偏态分布	题库 1	18.583	0.987	0	27.533	14.60%	0.947
	题库 2	15.523	0.991	0	22.067	19.98%	0.960
	题库 3	14.443	0.982	0	20.333	54.74%	0.966
负偏态分布	题库 1	12.688	0.920	0	18.100	14.37%	0.960
	题库 2	12.650	0.955	0	18.233	32.03%	0.965
	题库 3	11.547	0.908	0	16.567	63.88%	0.968

通过上述两份表格，可以看出，在不同的被试分布形态下，我们所构建的题库（题库 1）依然具有较高的使用效率，特别体现在题库使用均匀性（χ^2）、测验重叠率（$\hat{r}$），以及未使用的项目（never used%）这三个指标上。虽然我们无法知道我们所构建的题库在真实测验情境中的性能，但是通过模拟实验的结果显示，不管被试的分布形态如何，通过我们的方法所构建的题库表现均较优。由此可以为推断该题库在现实测验情境下的表现提供参考

意见 3：研究方法中基于 CDA 的最佳测验设计模式采用的最大题量融合方法忽略了项目曝光的问题。考虑两种情况：情况 1 在区间 A，每个被试都测了 5 个项目，融合后期望项目是 5；情况 2 在区间 A，只有一个被试测了 5 个题目，其余被试只测了 1 个题目，融合后期望项目仍是 5。这两种情况下，题目曝光差别肯定是很大的，而作者提出的测验设计模式中并没有考虑。

回应：审稿专家所说情况 2 正是最大题量融合原则的不合理之处。在探索每种认知结构的最佳测验设计时，若采用最大题量融合会造成区间内的项目数量虚高，如正文中表 4 所示。鉴于此，本文并没有采用最大题量的融合方法，而是改为采用另外两种区间融合方法：*M*+*SD* 和 p_{90}（见正文中表 4 的上一段）。使用这两种原则进行融合时，就避免了情况 2 的出现。其次，本实验虽然未在构建题库时专门考虑项目曝光的问题，然而，实验结果发现，当我们使用基于不同认知结构的最佳测验设计模式构建题库时，在题库使用指标上，特别是项目使用均匀性、未使用的项目数量上均具有较大优势。这表明，通过不同认知结构的最佳测验设计模式得到的题库可以有效地解决题库中项目曝光不均匀，曝光不足的情况。

意见 4：实验一假设被试认知结构服从均匀分布，这种分布的假设是否会对测验设计模式产生影响？为何这样设定？

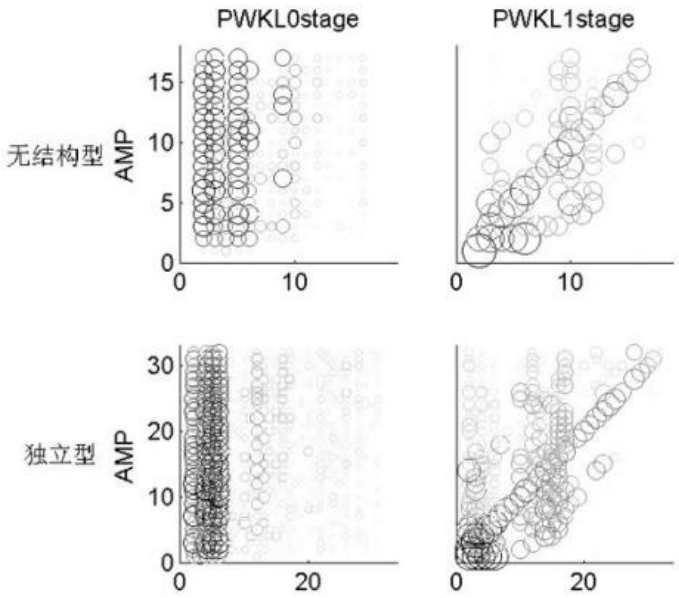
回应：采用被试认知结构服从均匀分布不会对测验设计模式产生影响。探讨每种认知结构的最佳测验设计模式，是对每种认知结构单独实施的模拟，不同认知结构之间互不影响。研究中为考虑被试分布的完备性，保证在测量目标领域范围内的每种认知结构都有，且每种认知结构的人数足够多，我们在实验一中设定掌握每种认知结构的被试各 100 人，通过对这 100 人在 CD-CAT 中所抽取的项目类型进行分析，得到每种认知结构的最佳测验设计模式。换言之，我们考虑的是能保证结果可靠性的被试样本设计，并没有从选取被试分布这一角度出发。与此同时，通过这种设计模式构造出来的被试分布确实是均匀分布。因此，出于以上考虑，采用被试认知结构服从均匀分布是合理的。此外，问题 1 的回应中我们补充的模拟实验结果也可从侧面佐证这一合理性。

意见 5：公式（2）下方，对于公式的解释中 s 和 g 使用的字体太怪，很难让对该模型不熟悉的读者认出。
回应：感谢审稿专家的提醒，已修改。

意见 6：图 1 中的第二和第三个图内容重复。因为只有 0 和 1 两个阶段，已知 0 阶段所占的百分比，就足以推知 1 阶段的百分比。
回应：感谢审稿专家的提醒，已修改。

意见 7：表五之前一段，作者之前已经提出了最大题量原则进行融合的各种问题，因而改用其他融合原则。为何在此又重新使用最大题量原则？
回应：这是我们疏于解释，感谢审稿专家指出这一问题，已在正文对应位置增加了简短说明。本质上，实验一可以分为两步，第一步主要探讨不同认知结构的最佳测验设计模式，第二步是在此基础上探讨如何构建能够满足所有认知结构类型施测要求的题库。在第一步中采用最大题量融合原则是不合适的，会导致区间内的项目个数虚高（即审稿专家在问题 2 中所述情况 2），文中也有对应说明。第二步采用最大题量融合原却是合适的，原因在于第二步需要考虑的情况和第一步有所不同。具体而言，假定有三种认知结构类型 A、B、C，经过第一步后，三者某一区间内所需的项目数量分别为 1、1、5。若按照第一步的融合方法 M+SD 和 p90，则必然会导致认知结构 C 在该区间上没有足够的测验项目，也就必然无法达到诊断该类型的最佳测验模式所需最低项目数量。此时，应当是使用最大题量原则进行融合，这样既满足了项目数量上“要求最严苛”的认知结构类型（C），也满足了其他的认知结构类型（A、B）达到最佳测验模式的项目数量要求。故使用最大融合原则是更为合理的。

意见 8：本研究中项目类型是按照考察属性个数规定的，同一项目类型包含了多种属性考察的方式。例如，同样是考察 3 个属性，在混合型中考察 A1, A3, A6 与考察 A2 A6 A5，题目的考察方式有所不同。考察 A1, A3, A6 相当于只考察 A6。
回应：本研究所得的研究结果均是基于每种典型项目考核模式。但是随着属性个数的增加，属性层级关系越松散，对应的每种典型项目考核模式越多，会导致结果呈现不够清晰。以 K=5 时无结构型和独立型为例。此时典型项目考核模式的种类分别为 16 和 31 种，若将横坐标换成每种典型项目考核模式，则横坐标上分别对应会有 16 和 31 个点，这种精确的分类一定程度上降低了结果的可读性，，如下图所示。因此，我们以项目考察的属性个数将项目类型分类画图，使得结果可以清晰呈现。但是本研究得到的研究结果均是基于每种典型项目考核模式上的探讨。



意见 9：表 5 之下第二段是对于表 5 中趋势的解释吗？希望作者能够在这一段解释的更清楚些，目前的叙述使读者很难理解。

回应：感谢审稿专家的建议，已修改。

意见 10：实验二只采用了独立性的属性层级关系，这样代表性够吗？我们都知道，独立型是一种非常简单的属性层级结构，在真实测验中比较少见。

回应：采用独立型的原因在于，CD-CAT 中现有的题库模拟方法均是基于独立型的，例如陈平提出的模拟题库的方法（Chen, Xin, Wang, & Chang, 2012; 陈平，2011；陈平，辛涛，2011a，2011b）以及 Cheng 的方法（Cheng, 2009, 2010; Zheng，Chang，2016;毛秀珍，辛涛，2013）。因此，为了可以有比较的基线，故选择了与其相同的属性层级关系进行实验。

……

审稿人 2 意见：本人认为作者提出的测验设计方法在理论基础和方法合理性上存在较大问题，使用起来比较麻烦，理由如下：

意见 1：CDCAT 的本质就是"因人施测"的测验，即已实现能够根据被试认知结构的不同测试不同的项目。这一点从作者的表 1 和表 2 结果也可以看出。只要题库基本覆盖所有的题目类型即可使用。作者提出的方法缺乏必要性。

回应：我们认为审稿专家提出“只要题库基本覆盖所有的题目类型即可使用”的说法似乎不太妥当。

理论上讲，认知诊断评价中的题库构建需要考虑以下三个方面：题库容量，Q 矩阵以及项目质量。首先，题库容量对于题库中项目的编写、呈现以及题库的发展、计划和执行（例如题库的更新）来说至关重要（Stocking，1994）。并且已有研究表明，基于定长的终止规则下，题库容量至少为测验长度的 12 倍（Stocking，1994）。其次，在传统测验形式中，Q 矩阵是认知诊断评价中的重要内容，其设计优劣会直接影响测验的效果（彭亚风，罗照盛，喻晓锋，高椿雷，李喻骏，2016）。在建设 CD-CAT 题库时，同样需要在属性这一纬度上认真审视题库的 Q 矩阵设计，但目前尚未有研究探讨该问题。在有关 CD-CAT 的研究中研究者提出了不同的题库模拟方法，主要有：陈平提出的模拟题库的方法（Chen, Xin, Wang, & Chang, 2012; 陈平，2011；陈平，辛涛，2011a，2011b）以及 Cheng 的方法（Cheng, 2009, 2010;

Zheng，Chang，2016;毛秀珍，辛涛，2013）。上述方法虽并未针对 Q 矩阵设计这一问题进行研究，但实际上也是不同的题库 Q 矩阵设计方式。最后项目质量方面，题库中项目质量越高越好。项目质量越高，意味着项目对被试在相应属性上掌握与未掌握情况的区分能力越好(Madison & Bradshaw, 2015)。

此外，若从实践的角度出发，题库建设需要耗费大量的时间、人力和财力（Wang, Chang, & Huebner, 2011）。题库是为了满足大型测验的需求，因此在项目数量和质量上都有较高的要求。若没有目的性地命制或选用项目，构建出来的题库质量是存疑的，而且会存在可预见的不必要的经济成本浪费。

通过实验二的实验结果表明，通过我们的研究方法得到的题库 Q 矩阵在题库使用上均优于陈平和 Cheng 的题库设计方法。并且依据实验结果，我们提出了在构建题库时，针对不同属性个数及其层级关系下题库容量和题库 Q 矩阵设计的一些建议。因此，本文提出的方法在题库建设方面具有一定的必要性。

意见 2：作者所提出的测验设计是需要依赖模拟 CD-CAT 并记录各个区间内的项目数量才能获得，这意味着只要属性及其层级关系发生变化后，都要经历模拟才能得到最佳测验设计模式，使用起来很不方便。而且只要题目参数、被试数量、题库等信息发生变化后，都需要重新模拟。这两点表明作者所提出方法的理论基础是不太合理的，即是说没有必要像作者研究的这样去构建测验。

Kuo, B. C., Pai, H. S., &de la Torre, J. (2016). Modified Cognitive Diagnostic Index and Modified Attribute- Level Discrimination Index for Test Construction, Applied Psychological Measurement, 40(5), 315-330.这篇文章提出了合理的、不依赖于 CDCAT 模拟方式的测验设计方法，请作者参考。

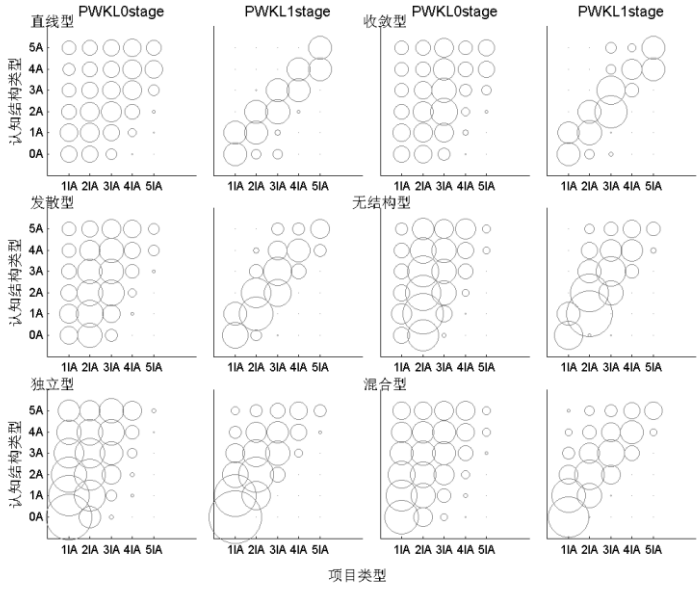
回应：首先，Ronald Flaugher（2000）指出要实现 CAT 的优势，题库中必须包含针对不同能力水平的高质量题目。同理，CD-CAT 可以为每种认知结构提供最匹配测验的前提是，题库容量足够大（每种典型项目考核模式出现的次数足够多），项目参数分布足够广。但这在受多种因素制约的现实情境下是很难完全实现的。因此，为了满足现实情境下题库的可操作性，本文借鉴 Reckase（2003, 2007, 2010）构建 CAT 题库的思路，使用 CD-CAT 的方法探索每种认知结构的最佳测验设计模式，在此基础上构建出了可以高效率使用的题库。采用五种基本的属性层级关系类型，此外为了仿真实际测验情境中可能存在较为复杂的属性层级关系模式，构建了混合型。通过实验，得到了在这六种属性层级关系下每种认知结构的最佳测验设计模式，并且不同属性层级关系下不同认知结构的最佳测验模式的设计具有共通性。因此，可以根据实际测验情景下的属性及其层级关系类型按照本文提出的最佳测验模式的设计规律对对应的认知结构类型设计测验，不需要再次模拟。

其次，不同的认知结构具有不同的测验设计模式。这主要体现在，每种认知结构所需的测验长度不同，所需的项目类型也不同。Kuo，Pai，de la Torre(2009)中所使用的方法，需要事前设定好测验长度。但是在不知道每种认知结构所需的最佳测验长度情况下，该方法并不能很好地使用。

意见 3：在编制测验时，通常无法知道题目的参数，而题目质量会严重影响 CDCAT 的表现，但作者在模拟 CDCAT 时需要依赖已知的题目参数，所以这种测验编制的指导作用不大。

回应：正如审稿专家所说，在编制测验时，通常无法知道项目参数。项目参数（体现着项目质量）是在项目编制完成并施测以后通过数据分析得到。必须先有测验项目，才能有项目参数，二者在逻辑上有明确的时间先后关系，这一点我们也是非常明确的。在认知诊断测验中，项目质量是关键参数，而属性则是另一个核心内容，同时也是我们本研究关注的焦点。项目

的好坏与否取决于它能否将不同认知结构区分开，这既涉及到项目质量，也涉及到项目考察的属性。虽然项目质量会影响 CD-CAT 中项目的选取，但是项目所考察的属性模式同样也会有影响（陈平，李珍，辛涛，2011）。我们研究的初衷在于通过探讨每种认知结构的最佳测验设计模式，构建目标领域内有针对性的题库蓝图，以优化对被试群体的诊断。依据这个蓝图，测验编制者便可有目标地编制高质量的测验项目入库。因此本研究本质上试图回答的是如何“先有测验项目”的问题，并把问题的关注点聚焦在属性这一维度上。在无法预知项目质量的情况下，需要考虑的是项目考察的属性如何才能有效的区分不同认知结构，即项目类型对于不同认知结构的区分能力。所以，本研究将题目参数作为已知条件通过模拟的形式给出。并且，通过逻辑分析我们认为项目质量的高低不会影响项目类型的选择，只会使得每种项目类型所需的项目个数稍加改变。为了谨慎起见，我们增加了一个模拟实验，探讨不同项目质量（U（0.15，0.35））下，不同认知结构在 0、1 阶段选出来的项目类型及其个数。实验结果如下图所示。

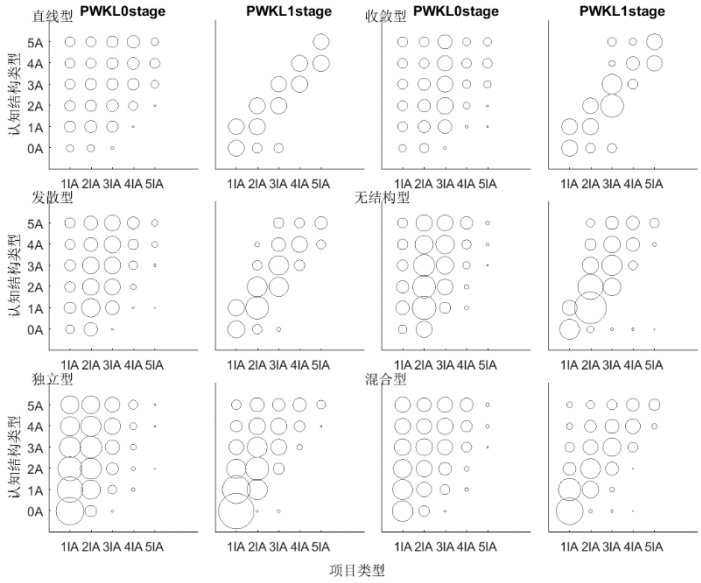


六种属性层级关系下所有认知结构类型在 0、1 阶段下选出的项目类型及其个数（s 和 g 服从均分分布 U（0.05, 0.25））

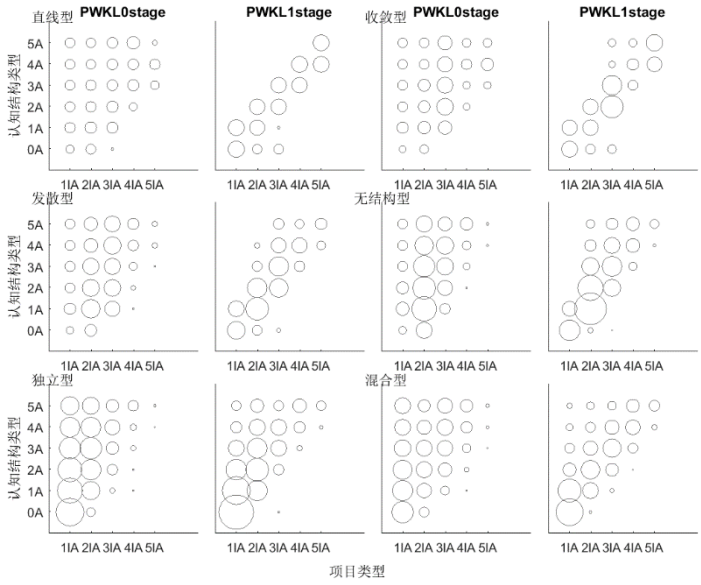
与正文中的图 2 对比可以发现，项目质量并不会影响选题策略所选的项目类型，整体模式并未发生改变。只是由于项目质量变差，为达到预先设定的终止规则，在所需的每种类型上的项目数量有所增加。因此，可以认为本方法对测验编制、题库建设具有指导作用，能够提供关于所需项目类型的重要参考信息，使得实践者能明确究竟需要考察那些属性的项目。同时可以根据实际项目质量的情况对项目类型的数量进行适当增减（在正文结论部分增加了相关说明）。

意见 4：第三步“融合”区间中的期望项目数量显然会受到被试量的影响。但在测验施测之前无法知晓会有多少人来做，因此，这个数量是没有意义的，所以 $M+SD$ 和 p_9o 也就没有意义，进而研究结果就不具有推广价值。

回应：通过逻辑分析，我们认为被试人数对研究结论应该不会产生实质影响。为了谨慎起见，我们根据专家意见，为了探讨每个区间内的期望项目数量受到被试数量的影响，我们增加了两个模拟实验：探讨当每种认知结构下的人数分别为 50 人、200 人时，在每个区间上的项目数量。结果如下图所示：



（a）六种属性层级关系下所有认知结构类型在 0、1 阶段下选出的项目类型及其个数（每种认知结构下 50 人）



（b）六种属性层级关系下所有认知结构类型在 0、1 阶段下选出的项目类型及其个数（每种认知结构下 200 人）

由实验结果可以看出，在每个区间内的期望项目数量不会受到被试人数的实质影响。因此，本研究提出的 $M+SD$ 和 p_{90} 是有意义的，且本研究的研究方法和研究结果具有推广价值。

意见 5：按照表 1 和表 2 中的举例来看，1 阶段被试的 KS 没有变化，是否所有被试都这样？是否意味着被试达到 1 阶段后就可以立刻终止测验？这样能够节省更多题量。如果不是，作者应该更换一个认知结构在随时变化的例子，这样不会被混淆。

回应：所有被试在进行 CD-CAT 的过程中都会经历这两个阶段，并且通过划分两个阶段，我们得到了六种属性层级关系下 0 阶段测验长度占总长度的百分比（请见正文中的图 1）。但这不意味着测验可就此终止。原因如下：因为 1 阶段才是严格按“因人施测”思想施测的阶段。当被试结束 0 阶段，此时所获得的被试认知结构估计值对应的掌握概率是处在较低的水平，误差也相对较大。此后在 1 阶段便依据被试当前认知结构估计值，从题库中选择项目施

测，从而能不断给被试认知结构的诊断过程提供更加丰富的信息，降低误差的影响，提高测验的信度。最终使得对于被试认知结构估计的精度达到预设的水平。因此，为了保证诊断结果的可靠性（信度，信息量），减少随机误差，并不建议在作答完 0 阶段的项目后就终止测验。

意见 6：表 4 想要表达的信息不清楚，本人不太理解。如果换其他的认知结构或区间，结果还会一样吗？

回应：是我们表述不清晰，感谢审稿专家指出，已对正文做出修改。

正文中表 4 是为了举例论证使用最大题量原则的不合理性。表 4 表明直线型情况下 1 阶段时认知结构为[0 0 0 0 0]的 100 名被试在在区间[1 0 0 0 0]上各自抽取的项目个数的频次分布。通过表 4 发现，在区间[1 0 0 0 0]上抽取了 2 个项目的有 86 人次，抽取了 3 个项目的有 6 人次，项目的有 9 人次，抽取了 6 个项目的仅有 1 人次，此时最大题量为 6。若按照最大题量的融合原则，则[0 0 0 0 0]的认知结构在区间[1 0 0 0 0]上需要 6 个项目，这会造成这个区间内的项目数量虚高，增加命题成本。如果换其他的认知结构或区间，在项目数量的频次分布上同样会出现较大的人数差异。例如，直线型情况下，认知结构为[1 1 1 0 0]的 100 名被试 1 阶段时在区间[1 1 1 0 0]上抽取的项目个数的频次分布如下表所示。可以看出，在区间[1 1 1 0 0]上，抽取了 1 个项目的有 14 人次，抽取了 2 个项目的有 75 人次，抽取了 3 个项目的有 1 人次，抽取了 4 个项目的有 6 人次，抽取了 5 个项目的仅有 1 人次。

认知结构为[1 1 1 0 0]所有被试在区间[1 1 1 0 0]上抽取的项目数量的频次分布

项目数量	人数
0	3
1	14
2	75
3	1
4	6
5	1

意见 7：作者在前言中提出了 CDCAT 的曝光控制问题，但在研究二中，作者并没有加入曝光控制。当加入曝光控制后，三个不同题库之间的题库使用情况以及平均测验长度可能会不一样，就有可能得到不一样的结论，即是说根据作者的方法构建的题库不一定是表现最好的了。

回应：首先，我们在前言中提到正是由于当前 CD-CAT 的题库在构建之时并未考虑不同认知结构对不同项目类型的针对性需求，所以导致题库利用上出现了曝光问题。正如实验二中题库 2 和题库 3 的实验结果所示。当我们使用基于不同认知结构的最佳测验设计模式构建得到的题库在题库使用指标上，特别是项目使用均匀性、未使用的项目数量上均具有较大优势。这表明，通过不同认知结构的最佳测验设计模式得到的题库可以有效的解决题库中项目曝光不均匀，曝光不足的情况。

其次，通过 Wang, Chang, Huebner (2011)和郭磊，郑蝉金，边玉芳（2015）的实验结果可以看出，项目曝光控制与判准率之间均在一个权衡问题。若在 CD-CAT 中加入项目曝光会在一定程度上降低对被试的分类准确性。而通过我们的方法所构建的题库，在不加入曝光控制的情况下，仍然可以在达到较高测量精度的同时，提高题库的使用效率。

意见 8：最后结论部分，作者给出的测验设计模式描述很不清晰，看起来很乱，本人看完之

后，还是不能掌握该如何编制测验。
回应：感谢审稿专家指出，已修改。

第二轮

审稿人 1 意见：感谢作者对于问题的仔细回答。经过修改，本文清晰许多，应该会为本领域其他研究者提供借鉴。
回应：感谢审稿专家对本文的肯定。

审稿人 2 意见：测验编制是学习评估中的一个重要环节，该研究探讨了不同认知结构下如何经济有效地设计测验，研究具有一定的现实意义。作者基本回答了审稿人的一审意见，并补充了一些实验作为依据，帮助审稿人进一步理解方法的优势。经过作者的修改，论文质量有了较大提升。但还有一些要注意的地方，供作者思考：
意见 1：需要进一步回答审稿人 2 的第 5 问。审稿人很好奇作者所举两个被试的例子：只要被试从阶段 0 进入阶段 1，之后的认知结构就不会发生变化了，如表 1 中第 4 列第 5 行至第 9 行，表 2 中第 4 列第 5 行至第 9 行。所有的被试都是这样的变化模式吗？如果存在这种规律，后验概率是否增加其实就无关紧要了。如果不是，建议作者更换被试 *n* 的举例，否则很容易产生疑惑。
回应：与 CAT 的施测过程会分为试验性探查阶段和精确估计真值阶段一样，所有被试在 CD-CAT 的作答过程都会经历 0、1 阶段，只是具体到每个被试时 0、1 阶段所需的项目个数不尽相同。试分析被试每做完一个项目后认知结构估计值可以看出，被试在进入 1 阶段之后认知结构估计值不会再有变化，但是其后验概率却在不断地增加；但是在 0 阶段时，其认知结构的估计会有波动（例如出现估计正确但随后又估计错误的情况），为此我们更换了被试 n 的举例，具体请见正文中表 2。此前文中对 0 阶段的解释，忽略了被试认知结构估计会出现波动的情况，现已修改对应的表述，具体详见表 3 下一段。

在 CAT 的试验性探查阶段会得到关于被试能力的粗略估计值，在随后的精确估计阶段，通过继续施测项目累增信息量以不断修正所得估计值，使其愈益接近真值。由于被试潜在特质在 IRT 中表现为连续值，而 CD 中是离散值。因此，在 CD-CAT 中被试认知结构估计过程的变化模式主要依据其后验概率的变化：通过 0 阶段的粗估初步得到被试认知结构类型，但其后验概率相对偏低，1 阶段时继续施测项目不断增加被试在该认知结构类型上的后验概率，以达到测验的测量精度要求。

意见 2：实验二的题库 1 和题库 2 容量为 152 题，而题库 3 容量为 360 题，题库数量相差两倍多，这样的比较是否公平？这会导致题库 3 的结果中，特别是χ2 值和 never used%（表 6 和表 7）很大，这很可能与题库 3 本身容量大有关系，请作者给予解释。该问题的回答也影响着 5.2 部分的结论：“基于不同认知结构的最佳测验设计模式构建出的题库，其使用效率比研究者常用的题库更高……..缓解了题库中项目浪费的情况”。
回应：实验二的目的在于比较不同的题库构建方法对题库使用效率的影响，这一比较是建立在被试容量、选题方法、终止规则等条件均相同的基础上。再者，题库 1 是基于不同认知结构的最佳测验设计模式构建得到，题量少是其相比与题库 3 的优势所在，并且从实验结果也得到了印证：在判准率相当的情况下，题库 1 的使用效率最高，题库 2 次之，题库 3 最低。这更进一步说明了本文 5.2 部分的结论，即基于不同认知结构的最佳测验设计模式构建出的

题库缓解了题库中存在的项目浪费情况。综上，我们认为在同样的实验条件下，将题库 1 和题库 3 进行对比分析，是合乎逻辑的研究选择。

意见 3：作者使用本文提出的方法设计测验，其中一个目的是要节约命题成本，可进一步理解为在保证估计精度的情况下，减少题量。尽管 CD-CAT 是低风险测验，但如果题库数量过低（如题库 1 和 2 中只有 152 题），还是会由于一些因素产生测验“公平性”问题，直接导致被试的认知结构估计发生偏差。请作者在讨论部分加上对该方面的理解。
回应：对于审稿专家所提出的测验“公平性”问题，我们的理解是，由于未能有效控制题库中高曝光率的项目，使得考生之间分享高曝光率的项目，从而影响其作答，导致被试认知结构估计发生偏差，进而引发“公平性”问题产生。造成项目过度曝光的可能原因主要有两方面：题库容量和选题策略。在选题策略一定的情况下，题库容量越小，项目过度曝光的可能性越大；在题库容量一定的情况下，不同的选题策略会有不同的项目曝光结果。

意见 4：作者在 5.1 部分是以 5 个属性的情况为例，差数了每种认知结构需要的考核项目数量。但实际测验中，属性数量是不尽相同的，请问其他属性数量时的结果也是如此吗？
回应：正如审稿专家所言，在实际测验中属性数量不尽相同。为此，我们在实验一中考虑了两种属性个数水平：5 个和 6 个，并且从两种属性水平的实验结果推导出了适用于任何属性个数下每种认知结构所需的项目类型及其数量的一般规律，正如 5.1 部分所示。非常抱歉的是，之前 5.1 部分存在笔误，误将“KA”写成“5A”，造成了审稿专家对于 5.1 部分仅讨论了 5 个属性的情况的误解，现已更正修改。

意见 5：同样是讨论部分，审稿人认为在讨论 5.2 部分时，不能仅仅依靠数据化的研究结果给予实践者建议，例如：“直线型、收敛型和发散型下所需的题库容量是对应的典型项目考核模式种类的 4-5 倍，无结构时为 3-4 倍，独立型为 2-3 倍”，更要结合实际中学科所考察知识内容板块、测验时间、测验长度等进行设计。
回应：诚然，不能仅仅依靠数据化的研究结果给予实践者建议。漆书青，戴海琦，丁数良（2002）指出，建设题库是一项系统工程，需要多学科专业人员（学科专家、心理与教育测量人员、计算机技术人员等）协同攻关，在科学的题库建设理论指导下有步骤地进行。我们的研究从理论上探讨了题库建设的一般框架，为科学建设题库提供新的可行方法。在建设题库的过程中需要切实考虑审稿专家所提到的多个因素。对此，实践者可以按照实际测验情境，使用本文提出的题库建设方法，事前通过模拟得到一个题库建设的初步框架，帮助实践者在构建题库时找到一个相对清晰的方向。在此基础上，再根据实际需要结合考察学科内容、测验时间等因素具体问题具体分析，进一步细化题库建设方案是较为实际的操作方案。

第三轮

审稿人 2 意见：
该论文研究角度新颖，为不同的认知结构被试提供了最佳的测验设计方案，可在一定程度上为实际题库建设提供借鉴和指导。经过两轮修改后，作者较好地回答了审稿人提出的审稿意见，基本达到了心理学报发表要求，建议发表。

编委意见：
该研究探讨了在不同认知结构下如何经济有效地设计测验，虽然只是一个模拟研究，离开实

际应用还有一定距离，但毕竟是对这个问题进行了初步研究，具有潜在的应用价值。 作者采用的研究方法正确，所得到的结果可靠，虽然考虑的因素还不够全面，例如没有充分考虑项目曝光率的问题，但还是对今后的研究具有很好启发作用。 作者对于审稿专家提出的问题都进行了较好的回答，还补充了一些实验。 该研究具有一定的创新性。

主编意见：English abstract attached and other minor editing in the article. Please consider.