认知诊断计算机自适应测验中选题策略的新进展

辛 涛 刘 拓*

[摘 要] 本文主要从提高测量精确性、控制题目曝光率、平衡测验内容三个方面总结了认知诊断计算机自适应测验的选题策略。文章首先简单回顾以往提出的选题策略,分析这些选题策略的不足,再分别介绍了三种改进方法。它们分别是毛秀珍提出的改善 PWKL 方法精确性的平均 PWKL 方法,陈平提出的控制曝光率方法,崇伟峰提出的通过后验概率到 0.5 的距离来平衡属性测量程度的方法。将三种方法与以往的选题策略比较,结果均显示三种新的选题策略效果更好。最后,作者探讨了认知诊断计算机自适应测验的选题策略中这三方面的平衡问题,以及选题策略以后的发展方向。

[关键词] 认知诊断计算机自适应测验;选题策略;曝光率;测验内容平衡

一、引言

心理学和教育学研究的不断深入和细化,对心理和教育测量也提出了新的要求。一方面,传统的项目反应理论(Item Response Theory,IRT)提供的关于被试心理特质或能力水平(θ)信息已经无法满足测验使用者的需求,他们希望更进一步的了解被试的认知状态(属性掌握模式)信息,获得更全面的诊断信息,从而帮助教学与干预,由此发展出了认知诊断理论(Cognitive Diagnostic Theory,CDT)。另一方面,传统的纸笔测验在大规模测验中费时费力,而人们希望测验时间更短,并且测验的精确性更高,由此发展出了计算机自适应测验(Computerized Adaptive Testing,CAT)。CDT 与 CAT 的结合则形成了认知诊断计算机自适应测验(Cognitive Diagnostic Computerized Adaptive Testing,CD - CAT),它可以提供关于被试的能力水平信息和属性诊断信息,同时也可以提高测验的效率和准确性(McGlohen & Chang,2008;辛涛,乐美玲,张佳慧,2012)。

与传统的 CAT 一样要实现 CD - CAT 的简短与高效,关键的环节就在于选题策略。无论是传统的 CAT,还是 CD - CAT 选题策略都可以分为两大类,一类是提高测量精确性的选题策略,另一类是非统计约束类的选题策略(Cheng & Chang,2009;Deng,Ansley,& Chang,2010;van der Linden,2000;毛

* 辛涛,教育部基础教育质量监测中心副主任,北京师范大学心理学院教授、博士生导师;刘拓,北京师范大学心理学院博士生,100875。本文为国家自然科学基金项目(31371047)阶段性研究成果。

秀珍,辛涛,2011a)。非统计约束类的选题策略又主要包括控制题目曝光的指标和平衡测验内容的指标。下文将从这两个类别三个方面出发分别介绍国内外已有的 CD－CAT 选题策略以及我们在此基础上进行的发展。

二、主要 CD－CAT 选题策略

在这两大类别的选题策略中,第一类选题策略主要关注提高测量的准确性,这是任何测验的首要目标。因此,在传统 CAT 中这一类的选题策略最多发展也最早,CD－CAT 的选题策略是传统 CAT 的扩展,也延续了这一特点,不过由于 CD－CAT 将被试能力区分成了离散的属性,有一些传统 CAT 中使用的选题策略不能直接推广到 CD－CAT 中(如 Fisher 信息量)。比较常用的用于提高测量精确性的 CD－CAT 选题策略主要有香农熵(Shannon Entropy,SHE)方法和 KL 信息量(Kullback-Leibler Information)类的方法(Xu,Chang,& Douglas,2003)。SHE 是一种反映随机变量不确定程度的指标,由 Shannon(1948)提出,后 Xu 等人(2003)将 SHE 方法引入 CD－CAT 的选题中。原本 SHE 的定义是随机变量分布概率的函数,随机变量概率分布越集中,SHE 的值越小,随机变量概率分布越分散则 SHE 越大。在 CD－CAT 的选题时,可将被试的认知状态属于不同属性掌握模式的概率分布看做随机变量的概率分布,被试认知状态属于不同属性掌握模式的概率差异越小,SHE 越大,被试认知状态属于不同属性掌握模式的概率差异越大,SHE 越小,也就越有把握把将被试归入某种属性掌握模式当中。KL 信息量类的选题指标包括 KL 信息量、PWKL(Posterior-Weighted KL,PWKL)信息量和 HKL(Hybrid KL,HKL)信息量,它们都是以 KL 信息量为基础的指标。其中 KL 信息量是在传统 CAT 中使用较多的一种信息量,Xu 等(2003)将其引入到 CD－CAT 的选题中。KL 信息量的构建是希望通过被试的实际作答反应来证明被试的认知状态是估计出的属性掌握模式而不是其他属性掌握模式。KL 信息量越大时,则说明越能有力证明被试的认知状态属于估计出的属性掌握模式。在 CD－CAT 中使用 KL 信息量进行选题和估计时,实际上存在着一个假设,即每一个被试在每一步 CD－CAT 的选题过程中,所有可能的属性掌握模式被考虑为被试真实认知状态的可能性是相当的。这实际上增加了选题的负担,而在实际测评中是存在先前的信息可以利用的,如一般认为本年度和上一年度参加测验的被试总体分布情况应该是相当的。因此 Cheng(2009)提出可以利用这样的先验信息对 KL 信息量进行加权,得到 PWKL 信息量。另一方面,根据 Henson 和 Douglas(2005),Henson、Roussos 和 Douglas(2008)的研究,题目对属性的区分是有差异的,有些题目可以更好的区分相似的属性掌握模式。Cheng(2009)提议使用属性掌握模式的欧氏距离来反映属性掌握模式的相似程度,欧氏距离越短说明属性掌握模式越相似。然后使用欧氏距离的倒数对 PWKL 信息量再次加权,就可得到 HKL 信息量。因此,HKL 信息量可以选出能够更能区分相似属性掌握模式的题目。

随着测量理论的发展,人们对 CAT 的要求也越来越高,第二类的选题策略也逐步受到研究者的重视。一方面,只考虑测量精确性的情况下经常会造成某类试题大量出现,如按信息量选题,则会造成高信息量的题目曝光率过高。而过高的曝光率不利于题库的安全,也会造成一部分试题的浪费。另一方面,不同于传统的纸笔测验,在 CAT 的过程中,由于每一位被试受测不同的试题,因此测试题的可比性和测验结果的公平性可能存在问题。为了控制题目曝光率研究者们在传统 CAT 的框架下发展出了很多方法,如随机化的选策略、分层的选题策略(Chang & Ying,1999)、SH 选题策略(Sympson & Hetter,1985)等等。但关于 CD－CAT 的框架下控制题目曝光率的选题策略研究还非常少,已发表的文献仅有 Wang、Chang 和 Huebner(2011)提出了两种限制性的随机化方法——限制进程法(Restrictive Progressive Method,RPM)和限制

阈限法(Restrictive Threshold Method,RTM)。为了平衡测量内容,在传统 CAT 框架下研究者们提出了一些方法,这些方法大体可分为两种:一是探索式算法,二是数学规划方法(毛秀珍,2012)。探索式算法一般是将一些限制规则放入选题方法或在选题方法中添加不同的权重来控制选题,而数学规划方法则往往是先根据目标使用数学算法进行组卷,再从新组合的测验中选题施测。CD－CAT 中平衡测量内容的选题策略有 Cheng(2010)的修正的最大全局区分度指标(Modified Maximum Global Discrimination Index,MMGDI)和潘奕娆(2011)结合认知诊断区分度指标(Cognitive Diagnostic Discrimination Index,CDI)后对其改进提出的 MGCDI。这两个指标都延续了探索式算法中加权的思想。在它们的权重中包含了属性目标测量次数和已测量次数的信息,如果某属性已测量的次数相对于目标测量次数过少,那么权重较大,则倾向于选择测量该属性的题目。

总之,提高测量精确性类的选题策略研究相对较多,但这些方法同样存在问题,而对于 CD－CAT 中如何控制曝光率和平衡测量内容的研究仍然处于初始阶段,有更多的问题亟待探讨。近年来,我们针对这三方面存在的问题分别进行了一些研究,提出了一些新的选题策略。

三、CD－CAT 选题策略的新发展

(一) 新的平均 PWKL 方法

在诸多提高测量准确性的方法中,PWKL 的表现较好,Cheng(2009)的模拟研究结果显示,无论在被试属性的判准率和模式的判准率上,PWKL 和 HKL 的效果都要优于 SHE 方法和 KL 方法,而 PWKL 和 HKL 之间的差异不明显。但 PWKL 也存在问题,在 CD－CAT 的早期因为所测题目过少,认知状态的估计是不准确的,有必要在选题策略中考虑这个影响。我们借鉴 Lima Passos 等人(2007)的思想,在 CD－CAT 初期利用认知状态后验概率的大小来选用多种认知状态来代替一种认知状态估计值作为被试属性掌握模式的估计值。随着 CD－CAT 的进行再使用这多种认知状态所对应的 PWKL 的算数平均数(APWKL)和几何平均数(GPWKL)作为选题方法(毛秀珍,2012)。

为了考察 APWKL 和 GPWKL 的表现,首先根据已有测验建立一个包含 321 题的数学能力的 CD－CAT 题库,题库涵盖 9 个属性。然后使用 DINA 模型模拟不同属性掌握模式的被试 1000 人。分别考察 KL、PWKL、HKL、APWKL、GPWKL 五种选题策略,在 15 道题、25 道题的 CD－CAT 情境下,属性判准率和模式判准率的情况。结果如表 1 所示。

从表中结果可见,在 20 道题目时,无论是属性判准率还是模式判准率,APWKL、GPWKL 和 PWKL、HKL 的差异都不大,但都优于 KL 信息量。在 15 道题目时,APWKL、GPWKL 的属性判准率略微优于 PWKL、HKL,同样这四种方法都优于 KL 方法。而模式判准率上 APWKL、GPWKL 要比 PWKL、HKL 高 0.09,比 KL 高 0.4。此结果也与我们(毛秀珍,辛涛,2011b)的模拟研究结果相同,两种新的平均 PWKL 指标可以有效提高判准率,特别是模式判准率,并且当测验长度越短时效果越好。

(二) 新的曝光率控制方法

为了了解 CD－CAT 选题方法对题目曝光率的控制,我们(陈平,李珍,辛涛,2011)基于 DINA 模型对 KL、PWKL、HKL 和 SHE 四种选题策略下题库使用的均匀性情况作了比较,发现效果均不理想。用图示法进一步分析 SHE 还发现,SHE 倾向于选择“g＋s”较小的题目。因此,我们提出了三种对选题曝光率进行控制的方法(陈平,2011)。方法一,由于 SHE 倾向于“g＋s”较小的题目,因此可以对“g＋s”进行排序,将题目分成 k 个水平,再使用 SHE 依次从每个水平中选题;方法二,将 Cheng 和 Chang(2009)提出的最大优先指标(Maximum Priority Index,MPI)引入到 CD－CAT 中,对题目的曝光率进行控制;方法三,就是方法一和方法二的结合,在方

法一分出水平后，每个水平的选题使用 MPI 进行。使用20道题的 CD－CAT，考察三种方法对题目曝光率的控制效果（见表2）。

表1　各选题方法属性判准率和模式判准率表

选题方法	测验长度	属性判准率									模式判准率
		1	2	3	4	5	6	7	8	9	
KL	15	0.97	0.83	0.99	0.99	0.94	0.79	0.99	0.79	0.99	0.52
PWKL	15	0.98	0.99	0.99	0.99	0.96	0.94	0.96	0.94	0.95	0.83
HKL	15	0.99	0.99	0.99	0.99	0.98	0.95	0.97	0.94	0.96	0.83
APWKL	15	1.00	1.00	0.99	0.99	0.99	0.98	0.99	0.98	0.99	0.92
GPWKL	15	1.00	0.99	0.99	0.99	0.99	0.98	0.99	0.98	0.99	0.92
KL	20	0.99	0.97	0.99	0.98	0.99	0.90	0.99	0.84	0.98	0.72
PWKL	20	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.97
HKL	20	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.97
APWKL	20	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.98
GPWKL	20	0.99	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.98

表2　20题的 CD－CAT 中各选题策略下曝光率控制情况表

选题策略	χ^2	测验重叠率	最大曝光率	最小曝光率	未用题数	曝光超过20%题数
SHE	146.7177	0.4626	1	0	180	32
方法一	139.1932	0.4416	1	0	143	23
方法二	0.0016	0.0546	0.0560	0.0550	0	0
方法三	0.0016	0.0546	0.0560	0.0550	0	0

表2中 χ^2 指标和测验重叠率都是评价测验使用均匀程度的指标，值越小越好，具体计算可参见陈平、李珍和辛涛（2011）的文章。从表中可以看出，方法一与 SHE 相比，题目曝光率的控制有所提升，未使用题数减少了37道，而曝光超过20%的题数减少了9道，χ^2 指标和测验重叠率分别减低到了139.1932和0.4416。而方法二和方法三相比 SHE 和方法一有了明显的提升，但它们之间没有差别。

（三）新的属性测量平衡方法

在 CD－CAT 中对测量内容的平衡，就是对属性测量程度的平衡，那么首先就需要定义如何来表征属性的测量程度。MMGDI 和 MGCDI 都是使用属性测量次数作为属性测量程度的表征，这样的表征方式并未触及到“诊断”的过程。所以我们认为可以用属性掌握的后验概率到0.5的距离来表征属性的测量程度（崇伟峰，2012）。该表征方法的思想是基于 Mislevy 和 Levy（2007）的以证据为中心的测验思想。这种思想认为新一代的测量就是以有限的观察反应（证据）来推断学生认知状态的过程。假设被试的认知状态只有“掌握”（0）和“未掌握”（1）两种取值，那么在没有任何被试反应信息的情况下进行猜测，猜对的概率就为0.5。随着测验的进行，就可以得到被试对于属性掌握情况的后验概率。将0.5看作一个基线标准，后验概率离0.5的距离越大，则通过后验概率推论被试是否掌握某属性的把握也越大，因此可以使用后验概率离0.5的距离作为属性测量程度的表征。

使用后验概率离0.5的距离就可以构建出一个作为权重的变量，将这个权重变量引入选题指标中就可以对属性测量内容进行平衡。如后验概率离0.5越大，则对被试的属性掌握情况判断越准确，而应更多的选择测量其它属性的题目。分别将到0.5的距离为0.1（DIS1），0.2（DIS2），0.3（DIS3）作为判断标准，也即当后验概率分别大于0.6，0.7，0.8或小于0.4，0.3，0.2时，则不再考虑对该属性的测量。在融合模型

的基础上比较这三种距离设定和使用 KL、PWKL 进行选题时属性测量的平衡性，模拟 755 道题，测量 5 个属性的题库，考察 32000 名被试的属性测量的均衡性，结果可见下表。

表 3　KL、PWKL 与新方法三种设定的属性判准率均衡性的比较

	KL	PWKL	DIS1	DIS2	DIS3
M	0.83	0.91	0.92	0.92	0.92
S	0.08	0.02	0.01	0.01	0.01
CV	9.64×10^{-2}	1.81×10^{-2}	1.81×10^{-2}	0.79×10^{-2}	0.84×10^{-2}

表中的 M、S、CV 分别表示属性判准率的平均值、标准差和变异系数。结果说明当加入后验概率到 0.5 的距离信息作为选题方法时，属性判准率的均值得到了提高，属性判准率之间的差异程度在减小。逐个考察每个属性的判准率情况也进一步证实了这一结果，DIS1、DIS2、DIS3 提高了在 KL 和 PWKL 方法下属性判准率较差的属性，即属性的测量更加均匀。

四、讨论

（一）选题策略三方面的平衡

对于一套 CD－CAT，首要前提就是能够准确测量被试，它体现在通过 CD－CAT 的施测能够正确测量被试的认知状态或属性掌握模式，这一前提直接决定了一套 CD－CAT 的测验是否有存在的价值。因此，与之对应，针对于提高测量准确性的选题策略的研究也是最多的。测验曝光率的问题，是一个题库安全性问题。若是一套低风险的 CD－CAT，则不需要太多考虑，而若是一套高风险的 CD－CAT 测验，过高的题目曝光率则可能导致测验的不公平和人力、物力的浪费。测量内容平衡性的问题，则实际是一个效度问题。CD－CAT 测验是否能均衡的覆盖到所需要测量的属性和潜在特质，关系到测验结果的比较与解释。

选题策略的这三个方面的关注角度不尽相同，但却存在着一定的联系。当关注测量准确性时，选题策略往往会倾向于选某种类型的题（如区分度较高的题目），则会造成某些试题的曝光率过高，而若采取随机化选题来控制曝光率则属性测量的准确性又较差（汪文义，2009；陈平等，2011）。当关注测量内容的平衡时，如使用后验概率到 0.5 的方法加权选题策略，由于平衡了每个属性的测量程度，所以使模式判准率有所提高，如图 1 反映了 KL、PWKL 两种选题策略和三种平衡属性测量的设定方法下，属性模式判准率的情况。可以看到平衡属性测量以后，模式判准率要比 KL 和 PWKL 方法要好。但另一方面，平衡属性测量程度，却依然面对题目曝光率的问题，每一个属性所涉及的题目中，总有一些会被经常选出，而另一些很少被选。继续考察五种方法下曝光率的情况，就可以发现平衡属性测量之后题库的使用均衡性变差（见表 4）。可见，如何在保证测量准确性的同时兼顾内容的测量，并且控制题目的曝光率，将三个方面所考虑内容进行平衡仍然有待探索。

表 4　五种选题策略下题目曝光率情况表

	KL	PWKL	DIS1	DIS2	DIS3
χ^2	157.27	194.71	217.92	261.07	264.01
曝光率 >0.2 的题目比率	0.01	0.01	0.01	0.01	0.02
曝光率 <0.002 的题目比率	0.74	0.79	0.81	0.83	0.83

（二）未来发展

选题策略在传统 CAT 中的发展已经相对成熟，但在 CD－CAT 却仍有许多问题需要解决。传统的 CAT 基于 IRT 的模型，而认知诊断的模型众多（DiBello & Stout，2007；余娜，辛涛，2009；甘媛源，余嘉元，2010；），并且构建基础差异很

大。因此,已经发展出的选题方法还十分有限,一方面需要开发更多新的选题方法,适应不同的认知诊断模型和 CD－CAT 情境。另一方面,已有 CD－CAT 选题方法的可推广性还需要更多的研究验证。

现有的 CD－CAT 的选题策略大多都聚焦在测量精确性的问题上,对于题目曝光率的控制研究还较少,而关注于平衡测量内容的选题方法就更少。所以发展非统计约束类的 CD－CAT 选题策略是一个重要的方向。特别是需要发展出能提高精确性,并同时控制曝光率和均衡测量内容的选题方法。

测验的发展使得施测者和受测者都对测验所能提供的信息提出了更高的要求,在 CD－CAT 的选题中则直接反映在对测量内容的控制上。需要控制测量的内容,首先又需要对于测量内容的进行表征,如 Cheng(2010)的方法以属性的测量次数表征,而我们的新方法则以后验概率到 0.5 的距离为表征。研究者们对此还未达成共识,从测验的结构上来表征,从属性测量的区分度来表征都是值得继续研究的方向。

最后,CDT 本身更加复杂,测验所测量的属性数量、属性粒度以及属性间的层级关系都会直接影响到认知诊断模型的估计和之后的诊断效果。这些因素的影响也需要逐渐引入到 CD－CAT选题策略的研究中。

参考文献:

陈平、李珍、辛涛,2011:《认知诊断计算机化自适应测验的题库使用均匀性初探》,《心理与行为研究》第 2 期。

DiBello, L. V. & W. Stout,2007, "IRT-based cognitive diagnostic models and related methods", *Journal of Educational Measurement*,44(4), pp. 285－291.

Chang, H. H. & Z. Ying,1999, "A-stratified multistage computerized adaptive testing", *Applied Psychological Measurement*,23(3), pp. 211－222.

Cheng, Y.,2009, "When cognitive diagnosis meets computerized adaptive testing: CD－CAT", *Psychometrika*, 74(4), pp. 619－632.

Cheng, Y.,2010, "Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method", *Educational and Psychological Measurement*,70(6), pp. 902－913.

Cheng, Y. & H. H. Chang,2009, "The maximum priority index method for severely constrained item selection in computerized adaptive testing", *British Journal of Mathematical and Statistical Psychology*,62, pp. 369－383.

Deng, H., T. Ansley & H. Chang,2010, "Stratified and maximum information item selection procedures in computer adaptive testing", *Journal of Educational Measurement*,47(2), pp. 202－226.

Henson, R. & J. Douglas,2005, "Test construction for cognitive diagnosis", *Applied Psychological Measurement*, 29(4), pp. 262－277.

Henson, R., L. Roussos, J. Douglas & X. He,2008, "Cognitive diagnostic attribute-level discrimination indices", *Applied Psychological Measurement*,32, pp. 275－288.

Lima, Passos, V., M. P. F. Berger & F. E. Tan,2007, "Test design optimization in CAT early stage with the Nominal Response Model", *Applied Psychological Measurement*,31, pp. 213－232.

McGlohen, M. & H. H. Chang,2008, "Combining computer adaptive testing technology with cognitively diagnostic assessment", *Behavior Research Methods*,40(3), pp. 808－821.

Mislevy, R. J. & R. Levy,2007, "Bayesian psychometric modeling from an evidence-centered design perspective", In C. R. Rao & S. Sinharay (eds.), *Handbook of Statistics*, *Volume* 26: *Psychometrics*, Amsterdam: North-Holland: Elsevier, pp. 839－866.

Shannon, C. E. ,1948, "A mathematical theory of communication", *The Bell System Technical Journal*,27, pp. 379 - 423.

Sympson, J. B. & R. D. Hetter,1985, "Controlling item-exposure rates in computerized adaptive testing", Paper presented in Proceedings of the 27th annual meeting of the Military Testing Association, San Diego CA: Navy Personnel Research and Development Centre, pp. 973 - 977.

van der Linden, W. J. ,2000, "Constrained adaptive testing with shadow tests", in W. J. van der Linden & C. A. W. Glas (eds.), *Computerized Adaptive Testing: Theory and Practice*, Norwell, MA: Kluwer, pp. 27 - 52.

Xu, X., H. Chang & J. Douglas,2003, "A simulation study to compare CAT strategies for cognitive diagnosis", Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

(责任编辑:蒋永华)

Progress of Item Selection Methods in Cognitive Diagnostic Computerized Adaptive Testing

XIN Tao, LIU Tuo

Abstract: In cognitive diagnostic computerized adaptive testing (CD - CAT), researchers have to take three aspects into consideration in item selection procedure: how to enhance accuracy of item selection methods; how to control the item exposure; and how to balance the attribute coverage. We reviewed some existing important item selection methods, and analyzed their drawbacks in this article. We also introduced three improved methods: Mao Xiuzhen's two averaged PWKL, Chen Ping's exposure control method, and Chong Weifeng's attribute balancing method which uses the distance between a posterior probability of attribute mastery and 0.5. Compared with the old methods, these three new methods can do better in item selection. In the end, we discussed the relationship between the above-mentioned three aspects and the research direction of item selection methods in CD - CAT.

Key words: cognitive diagnostic computerized adaptive testing; item selection; item exposure; balancing attribute coverage