

适应性测验的发展 :历史与现状

张青华¹ 袁一萍² 张厚粲¹

(1. 北京师范大学 心理学院 北京 100875 2. 江西师范大学 教育学院 南昌 330027)

摘 要 :适应性测验作为一种测验形式 ,与传统的纸笔测验形式相比 ,具有省时、高效等很多优越性。测验应该适合于被试的理念最初可见于比内的智力测验。上世纪七十年代以来 ,适应性测验的研究从经典测量理论阶段发展到项目反应理论阶段 ,经历了从二阶段、三阶段、多阶段测验、固定分支测验和分层适应性测验的发展 ,到现在的计算机化适应性测验研究。随着项目反应理论和计算机技术的发展 ,计算机化适应性测验已经在教育和心理测验领域中得到广泛应用。目前对它的研究日益深入 ,主要有项目克隆、项目曝光、多维适应性测验、被试诊断、人格适应性测验等问题。

关键词 适应性测验 纸笔测验 项目反应理论

中图分类号 :B841.7

文献标识码 :A

文章编号 :1003 - 5184(2006)02 - 0084 - 04

适应性测验(adaptive test)又可称为裁剪测验(tailored testing),它主动地适应被试能力水平,不同的被试做答不同的项目,随后施测的项目是根据被试对前面施测项目的反应来进行挑选。它的主要特点就是为每名被试挑选出适合其能力水平的一组测验项目。

1 适应性测验的发展历史

自从 1900 年以来,由于第一次世界大战的需要,传统的纸笔式心理和教育测验技术得到迅速发展。纸笔测验通常是不论被试水平的高低,每人都接受相同的测验项目。它无法精确地测量所有不同能力水平的被试。适应性测验的基本思想就是对每一名被试施测与其能力水平相当的项目,从而最节约、最有效地测量出被试的特质水平。

测验适应于被试能力水平的思想很早就已经存在,甚至比传统纸笔测验的发展还早。比内(A. Binet)在他最早编制的比内智力测验和后来出版的斯坦福比内智力测验中,就已经具有了适应性测验的思想。在比内智力测验中,不同年龄儿童在测验时测试项目的起始点不同、在施测过程中即刻进行评分、采用适应性的方法挑选测验项目(即由前一组施测项目上的表现决定后一组施测项目)。虽然早期的智力测验已具有适应性测验的思想,但它们是靠人工操作的适应性测验。二十世纪五十年代以后,美国军队研究者开始探索用测验机器进行适应性测

验的可能性。到六十年代后期,美国海军人员培训研究计划开始支持洛德(Frederic Lord)提出的项目反应理论研究和适应性测验研究^[1,2]。洛德对适应性测验的发展进行了开拓性的工作,他解决了团体测验的理论结构、个体的适应性测验以及许多实际的细节问题。在洛德对适应性测验进行研究的同一时期,美国海军支持了“计算机化适应性测验应用”的研究项目^[3]。二十世纪七十年代后期随着计算机技术的发展、项目反应理论的发展和应用,计算机化适应性测验的研究和应用也迅速展开。

从心理与教育测量学理论发展的角度来看,可以把适应性测验的发展划分为经典测量理论发展阶段和项目反应理论发展阶段。在经典测量理论发展阶段,适应性测验的施测策略经历了从二阶段测验、三阶段测验、多阶段测验、金字塔测验和分层适应性测验的发展历程。最早的适应性测验是出现在二十世纪五十年代后期,它用于测量能力,是由两阶段的纸笔测验(two-stage paper-and-pencil test)组成。被试在第一阶段完成由一组中等难度项目构成的测验,然后根据他们在第一阶段测验上的得分,来选择第二阶段的测验。如果他们在第一阶段测验上作答很好,则在第二阶段提供比第一阶段测验更难的一组测验项目;如果答得不好,则提供比前一阶段测验更容易的一组测验项目。由于能够提供高低难度不同的测验,所以这种测试能够对宽范围的能力水平

提供较好的测量精度。后来两阶段适应性测验又被扩展为三阶段、四阶段、甚至更多阶段的适应性测验。在适应性测验的每一个分支阶段,被试的行为决定着后一阶段测验是否比前一阶段测验更难或更易。这样在整个能力量尺很宽的范围内都能有较高的测量精度。这种多阶段的适应性测验,它每一阶段都由一组同质项目构成。在后来的适应性测验中,一个项目替代了一组同质项目,发展为单个项目的固定分支适应性测验,也就是说每一阶段施测的是一个项目而不是一组同质项目。测验项目的结构被建构成一个金字塔形,从塔顶开始,第一阶段的项目难度为中等水平,塔顶下面(第二阶段)有二个项目,与塔顶项目难度相比,一个较难,一个较容易。如果被试答对塔顶的项目,则施测较难的项目,否则,施测较容易的项目。对于随后的每一个阶段,它的每一个项目下面都有两个项目。对于这种金字塔形的适应性测验,它有 $2^n - 1$ 种项目组合方式(n 代表金字塔的阶段数)。随着 n 的增加,每一份测验会越来越适应于对每位被试能力的测量。

在固定分支适应性测验的基础上,后来发展了分层适应性测验。分层适应性测验就是在能力量尺上特别难至特别容易的范围内,能力水平被划分为由低到高的若干个层,每一层中项目的难度水平都反应了该层的能力水平,前一层项目的平均难度水平比后一层项目的平均难度水平要高。在进行施测时,根据被试先前的信息(如被试以前测验的成绩、自我报告等)来决定首先对被试施测哪个项目层。与以前的多阶段适应性测验、固定分支适应性测验相比,分层适应性测验具有两个显著不同之处。第一,对于每一名被试而言,每一层的测验起始点都可以不相同,即对每一名被试可以施测不相同的初始项目;第二,分层适应性测验采用了变化的测验中止规则。也就是说,对不同被试施测的项目数量不等,有长有短,因为所有施测的项目都更适应于特定被试。

上述适应性测验方法能够在宽范围内较好地测量被试的特质水平,但也存在着一定的局限性。首先,测验初始项目的挑选和测验的中止规则是武断的。根据这些规则,相同的被试可能会接受完全不同的项目组 and 得到不同的测验分数;其次,由于每名被试测的是不同项目组,而且每名被试的测验长度

也不相同,造成评分复杂,依据经典测量理论进行评分出现了问题,最后,这些适应性测验方法都受限于项目难度信息。因为挑选项目和施测都依据项目的难度,而在经典测量理论中项目难度水平依赖于被试群体的水平。

随着心理计量学的新发展,尤其是项目反应理论的建立与发展,极大地促进了教育和心理测验的发展。与经典测量理论相比,项目反应理论具有如下一些优越性:项目参数与特质参数(或能力参数)不变性;项目难度和被试特质水平被定义在同一个单位量尺上;项目信息函数和测验信息函数概念的提出;应用项目反应理论建构测验具有主动性、可控性等。也就是说,可以提前控制测验的难度、测量的误差,可以根据项目信息量来挑选项目^[4]。

项目反应理论所具有的这些优点为适应性测验的发展提供了更加坚实的基础,同时也解决了基于经典测量理论的适应性测验所面临的困难。第一,项目参数不变性使得建立所有项目都具有相同量尺单位的题库成为可能;特质参数不变性可以使测量不同项目的被试得分能够直接进行比较;第二,项目难度与特质水平处于相同的量尺,在适应性测验的施测过程中,可以直接根据估计的被试特质水平来寻找相应难度水平的项目;第三,项目信息量和测验信息量概念的提出,使得在适应性测验中,可以根据估计的被试特质水平,选择能够提供最大信息量的项目来继续施测,从而保证能够以最少的项目,高效准确地测量出被试的特质水平,从而克服了在经典测量理论中仅依据难度来挑选项目的局限;同时也可以预先设定适应性测验对所有被试特质水平测量误差的标准。

事实上,没有项目反应理论作基础,适应性测验不可能有如今的发展。同时,计算机技术的发展使得适应性测验的思想得到切实可行的实现。适应性测验的思想、项目反应理论的理论基础和计算机技术,这三者的结合造就了计算机化适应性测验的产生和发展。

2 当前计算机化适应性测验的应用与研究发展

2.1 当前计算机化适应性测验的应用

自从上世纪七十年代末计算机技术应用于适应性测验后,计算机化适应性测验的优越性很快受到重视,得到迅猛地发展,目前已经在一些大规模测验

中得到广泛应用。例如:美国的研究生资格考试(the Graduate Record Examination, GRE, Educational Testing Service)国家护士资格考试、国家医生资格考试。此外,美国国防部门的军队职业能力倾向成套测验(Armed Services Vocational Aptitude Battery, ASVAB)^[10]也采用了计算机化适应性测验形式。计算机化适应性测验不仅被应用于教育测量领域,而且也逐渐在心理测量领域中的人格测量、心理诊断^[11]等方面得到应用,如明尼苏达多相人格调查表—成人版(MMPI-A)的计算机化适应性测验。

2.2 计算机化适应性测验研究新方向

计算机技术及网络技术的迅猛发展,促进了计算机化适应性测验被广泛应用。与此同时,项目反应理论的发展,更多反应理论模型的开发,促使人们对计算机化适应性测验的研究也越来越深入。最初对计算机化适应性测验的研究主要围绕着它的几个基本方面:题库的建构、项目挑选策略、被试特质水平估计的方法、测验的中止规则等。例如,如何建构适应性测验的题库、在施测过程中根据何种规则为被试挑选下一个作答项目、采用哪种方法更能准确地估计被试的特质水平、以及采用什么样的测验的中止规则等。这些研究的成果对于计算机化适应性测验的不断完善起了重要作用。然而,实践中不断出现新问题,推动计算机化适应性测验的研究又出现了新的方向。

当前,主要研究有适应性测验的项目克隆、测验项目的曝光、多维适应性测验、被试认知诊断、人格适应性测验等几个方面^[5-8]。

适应性测验的项目克隆。克隆项目就是应用计算机程序根据已有的项目模板来生成新的项目。项目并不预先存在,而是在施测的过程中被编写。计算机化适应性测验需要建构题库,为了增加适应性测验题库中项目的数量,并且减少项目编写的费用,人们提出了使用项目克隆技术来生成适应性测验的项目。目前对适应性测验的项目克隆研究主要有:项目克隆后项目参数的变动、适应性测验施测时对克隆题库中的项目挑选等问题。

测验项目的曝光。测验项目的曝光问题是适应性测验所需要解决的一个重要问题。所谓项目的曝光是指在施测过程中,项目被选中而施测于被试。它不仅涉及到单个项目被施测的概率大小,同时也

涉及到题库中所有项目能否都被合理利用。对于适应性测验而言,如果题库中的项目曝光度太高,即项目被施测的概率太大,题库中的项目很容易被人们了解,这样题库的安全性将受到威胁,难以保证测验的有效性和公平性。在适应性测验中最常用的项目挑选策略是有限定的最大信息量项目挑选方法,这种策略虽然能够通过对被试施测最大信息量的项目,从而达到最好的测量精度,但是它通常使得具有高信息量的项目曝光率太高,而具有中等或较低信息量的项目没有被使用。为此研究者提出了许多解决方法,如 α 分层法(alpha-stratified method,题库中的项目按照区分度参数进行分层,施测时限定了从某一层中挑选项目)条件挑选策略(conditional selection strategies,根据设定的标准来控制项目被施测的概率)随机策略(randomization strategies,从接近最佳项目的项目组中随机挑选项目,而不是只挑选具有最大信息量的项目)。

多维适应性测验。多维适应性测验(Multidimensional adaptive testing, MAT),是指一个适应性测验测量几个潜在特质。通常人们把多维划分为两种:项目间多维(Between-Item Multidimensionality)和项目内多维(Within-Item Multidimensionality)。项目间多维测验是指测验测量多个特质,但其中的每个项目仅测量一个特质,如包含几个分测验的测验集,每个分测验测量一个特质;项目内多维测验是指包含多维项目(同一个项目测量多个潜在特质)的测验,例如作文题既能测量语言能力也能测量思维逻辑能力。目前人们比较关注多维适应性测验的应用和测量的有效性方面。

被试的认知诊断。认知诊断是当前测量领域中一个比较新的研究方向^[9]。对于被试而言,测验的功能不应仅是报告其能力水平(测验分数),而且更重要的是能对其能力结构进行诊断,教师可以根据测验的诊断结果来帮助被试的学习和知识结构的掌握。传统适应性测验的主要目标是精确、有效地估计出被试的能力水平。然而有时在实际的测试当中仅估计出被试的能力水平还不够,还需要对被试做出诊断(或分类),如通过或不通过、掌握或没掌握、被试具有怎样的能力结构等。随着人们对测验要求的不断提高,诊断型适应性测验的研究越来越得到人们的关注。

人格计算机化适应性测验。计算机化适应性测验最初在教育测量领域得到广泛的关注,后来人们开始对人格的计算机化适应性测验进行研究^[10]。目前对人格计算机化适应性测验的研究主要有几个方面:人格适应性测验的效度研究及其与纸笔人格测验的可比较性研究、人格适应性测验不同中止规则的比较、人格适应性测验中人格特质水平估计精确度的比较研究等。大多数研究表明,在不失测验精确度的情况下,人格适应性测验比相应的纸笔测验节省 50% 左右的项目,可以大大提高人格测验的效率。

参考文献

1

Lord F M. Some test theory for tailored testing. In W. H. Holtzman(Ed.), Computer – assisted instruction , testing , and guidance(pp.139 – 183). New York : Harper and Row ,1970.

2

Lord F M. Tailored testing , an approximation of stochastic approximation. Journal of the American Statistical Association , 1971 ,66 :707 – 711.

3

Weiss D J , Betz N E. Ability measurement :Conventional or adaptive ?(Research Report 73 – 1). Minneapolis : University of Minnesota ,Department of Psychology ,Psychometric Methods Program ,1973 ,February.

4

Hambleton R K , Swaminathan H. Item response theory : Principles and application. Boston : Klumer – Nujhoff publishing , 1985.

5

Glas C A , van der Linden W J. Computerized adaptive testing with item cloning. Applied Psychological Measurement 2003 ,27 (4) :247261.

6

Wang Wen – Chung , Chen Po – Hsi. Implementation and measurement efficiency of multidimensional computerized adaptive testing. Applied Psychological Measurement ,2004 ,28 :295 – 316.

7

Parshall C , Cristine Harmas , J. & Kromrey J D. Item exposure control in computerized adaptive test : the use of freezing to augment stratification. Florida Journal of Education Research , 2000 ,40(1) : 28 – 52.

8

Meijer R R , Nering M L. Computerized adaptive testing : Overview and introduction. Applied Psychological Measurement ,1999 ,23 :187 – 194.

9

Sands W A , Waters B K , McBride J R. Computerized adaptive testing : From inquiry to operation. Washington D. C. : American Psychological Association ,1997.

10

Butcher J N , Perry J , Hahn J. Computers in clinical assessment : Historical developments , Present status , and future challenges. Journal of Clinical Psychology , 2004 ,60(3) :331 – 345.

The Development of Adaptive Tests

Zhang Qinhua¹ ,Yuan Yiping² ,Zhang Houcan¹

(1. School of Psychology , Beijing Normal University , Beijing 100875 2. Education College ,Jiangxi Normal University ,Nanchang 330027)

Abstract : Comparing with traditional paper – and – pencil tests , adaptive tests have many advantages , such as saving time ,highly efficient , etc. The idea that the tests should adaptive examinee firstly came forth in the Binet Intelligence test . Since the early 1970s , the development of adaptive tests has been divided into two phrases from Classical Test Theory to Item Response Theory. And the research of adaptive tests has gone through from two – stage , three – stage , multi – stage , fixed – branching and stratified adaptive tests to computerized adaptive tests. With the development of Item Response Theory and the computer technology , adaptive tests are used more and more widely in the areas of educational measurement and psychological measurement. In the recent years , the emphasis of research in adaptive tests are on the directions of item cloning , item exposure , multi – dimensional computerized adaptive testing , diagnosis on examinees , personality computerized adaptive test , etc.

Key words : adaptive test ; paper – and – pencil test ; Item Response Theory