简论计算机自适应语言测试的工作机制*

张 宝 钧

提要 本文分别从题项反应理论(IRT)与计算机自适应语言测试(CALT)的关系、CALT的题库建设等方面探讨了CALT的工作机理及其优缺点,指出在大规模的语言测试中采用CALT是大势所趋,我国应及早开展对CALT的研究与开发。

关键词 计算机自适应语言测试;题项反应理论;题库

一 引言

计算机自适应测试(CAT, Computer-adapted Test)是网络计算机技术和教育测量理论相结合的产物,在各种资格考试中已得到了广泛的应用。语言测试领域对计算机自适应性测试的开发和效度研究也已经开始,其中最有影响的要数托福考试(TOEFL)。它于1998年7月就在美国以及少数其他国家进行了机考,2001年在全世界普及。国外其他的机考还包括伯明翰大学开发的法语、德语、西班牙语以及英语作为第二语言的分级考试,佐治亚州立大学开发的英语作为第二语言的听力考试,以及蒙特利尔大学开发的法语计算机自适应性水平测试(CAPT)等。目前国内的计算机自适应考试研究刚刚起步,对这方面的论述及研究较少,本文试图就计算机自适应考试的工作机理和相关的问题进行初步的探讨。

二 简要回顾

早在1968年,Green(1970)就预测说"测试必将被计算机征服"。20年后,Tung(1986:13)宣称,"由于心理测量理论的发展和计算机在教育系统中的使用越来越普及,使用高速计算机对知识和技能进行准确、有效和个人化测量的时代已经初露曙光"。从Green的预测到Tung这二十年间,正是计算机在教育测量领域应用的起步阶段。1975年,第一次专门探讨计算机自适应考试的大会在美国海军研究院和美国内务委员会的赞助下召开,之后又分别于1977年和1979年在明尼苏达大学召开了两次(Weiss 1983)。这三次大会的召开,极大地促进了计算机自适应测试技术、机考对考生的影响,促进了计算机自适应考试的效度等理论和应用方面的研究,使它的应用范围越来越广,技术也日臻完善和成熟。

在这一阶段,由于计算机强大的数据处理能力,使得先于其出现的题项反应理论(IRT, Item Response Theory,又称潜能理论 Latent Trait Theory)的实际应用有了物质基础,正是二者的结合才诞生了真正意义上的计算机自适应考试,进而为心理测量开辟了一块新天地,并对

* 本文的写作得到了国家基金委的资助。

以后相关领域的测验产生了重大的影响。

三　计算机自适应测试与题项反应理论

计算机自适应测试又被称为个性化的考试。与传统的纸笔考试不同的是，它使用计算机介质，键盘操作，这一点和计算机辅助测试(Computer-Assisted Test)相同，但在本质上却大相径庭。“传统考试中，每个考生，不论其能力高低，都必须做同一套试题，而自适应考试则可以根据考生的实际表现挑选出适合考生能力的试题，这与传统考试形成鲜明的对比。”(Canale and Baker 1984)计算机自适应测试的题项不是直接和线性的，即难易程度不是事先定好的，而是根据受试在前一个题项中所表现出来的能力或潜在能力而有针对性地选择出来的。换句话说，计算机随时都在对试题进行在线调整，以保证挑选出能最大限度测量出受试能力的题项。这就像量体裁衣一样，计算机在为受试“量身定做”考题，并按照这一程序，不断地对受试的能力做出判断并选择出测量这种能力的最佳题项，直到对受试的潜在能力做出全面的测量为止。因此，如果考生在最初的五个题项上表现欠佳，计算机会自动挑选更容易的试题。这很像跨栏比赛，假如我们的跨栏高度从 10cm 依次增加到 100cm，如果赛手比赛中把前五个栏都碰倒了，就没有必要让他再往前跑，把后面的五个栏也踢倒。同样，如果有些选手毫不费力地跨过第五个栏，那么，从他们可以跨过 50 厘米栏这一事实，可以肯定地推断出他们具有跨过 30 厘米和 40 厘米栏的潜在能力。在实际的测试中，难度太大的题项无法传达给我们有关受试的有用的信息，只能增加受试的焦虑程度。而自适应考试能以一种很友好的测试形式，根据考生的能力挑选试题，这就使受试在考试过程中非常放松，因而能更准确地测量出受试的能力。

题项反应理论实际上是一组不同的数学模型，这些模型描写了“对一个题项的回答和受试的潜在能力之间可能的关系”(Wiss 1983)。它包含下列几个要素。

(1)必须有一个刺激变量。这些变量可以是能力测试或成绩测试题项，也可以是个性调查问题。

(2)这些题目要由受试作出回答。

(3)为了让心理测量专家根据考生对刺激变量做出的回答推断出他的潜在能力，观测到的回答和其中隐藏的能力水平之间的假设关系是由一个能描写这种关系的等式来表现的。

因此，题项反应理论的核心就是这个能描写对刺激做出的反应与这个刺激变量所隐含的能力水平之间关系的等式。在这个等式中，刺激变量的特点是已知数，潜在能力模型要做的是根据受试对刺激变量做出的可观察到的回答估算出受试的能力水平。换言之，如果受试对一些具有已知特征的题项做出了回答，潜在能力模型就可以从这些回答中推断出受试的能力水平。

在题项反应理论模型中，最常用的是单参数、双参数和三参数逻辑模型。顾名思义，单参数模型是在一个单一的潜在能力——难度连续体上测量受试的表现；双参数模型在单参数模型的基础上增加了区分度；三参数模型除了能力——难度水平、区分度外，又增加了第三个参数，即猜测参数以及其他两个模型不包括的其他测量误差。由于复杂的模型需要的样本比较大，如三参数模型需要 2000 人以上的样本，双参数模型要求 400 人以上，因而最低只要求 200 人样本的单参数模型在计算机自适应测试的题库建设中较为常用，但它也存在明显的局限。

虽然题项反应理论在 50 年代就已成型(Lord 1952)，但一直局限于理论上的研究，没有投

人实际应用。主要原因就是缺少分析大型复杂数据的计算设备，而60年代末计算机的出现，特别是70、80年代功能强大的微型计算机的出现恰好满足了它对数据分析能力的要求，因而潜能理论具有了实用价值，测算潜能理论中数学模型的参数的程序具有了可操作性，因为计算机完成了两件传统方法无法完成的工作：(1)决定潜能数学模型中定义常量的测试题项的特点，(2)从受试对具备已知特征的题项的回答中测算出受试的能力水平。完成了这两件工作，潜能理论的实用价值以及相对传统测试理论所具有的优势就体现了出来。因此，潜能理论和计算机的结合是自然的，也是必然的。

和潜能理论一样，自适应性测试也是先于计算机出现的。第一个开发自适应性考试的是Alfed Binet(Weiss 1983)。他的测试策略虽然相对简单，但也具备了以题项反应理论为基础的计算机自适应测试的基本特征，但它对受试能力测量的准确度却无法与题项反应理论相提并论，而且由于没有计算机的辅助，它的普及使用几乎是不可能的。计算机投入使用后，出现了很多计算机辅助性测试，但除了测试实施的媒介由试卷改为计算机，手工书写改为键盘操作外，这些测试在理论上并没有实质性进步，基本还停留在传统的测试理论上。因此，只有当自适应性测试与计算机结合，以题项反应理论为基础进行题库建设和测试数据分析，才算形成了真正意义上的计算机自适应性测试，测试理论与实践才算有了实质意义上的变革。

可以说，没有计算机，题项反应理论只能作为没有实用价值的理论被束之高阁，是计算机以及后来的计算机自适应考试为其找到了用武之地，并使传统测试理论向前迈进了一大步。而计算机自适应性测试，虽然不依赖题项反应理论也可以存在，但题项反应理论的介入使计算机自适应测试的实施更加有效，测量也更加准确。换句话说，没有题项反应理论，也就没有真正意义上的计算机自适应测试，二者是相辅相成的。

四　计算机自适应测试的题库

计算机自适应测试作为一种计算机测试，需要很多的软硬件设备，主要包括电脑、在电脑上为受试在线选择、评判题项的程序以及一个可供程序选择的软件。其中和语言测试工作者最相关的也是最重要的就是题库。题库的好坏直接关系着评估结果的效度。

4.1 定义

一个题库就是一定数量、按照规定的内容说明和题项参数编写的、用于测量受试多层次能力的题项的集合。真正的题库必须具备如下五个特征(Henning 1991)。

(1)题库中的题项已经在受试总体中试用过、分析过，确实适合目标受试总体。实际的题项选择标准将根据试题编写方法、受试总体的特点以及测试的目的而变化。

(2)题库中的题项根据一定的系统被汇集在一起，这个系统能保证根据不同的测试目的即时检索并快速显示题项。

(3)题库中的题项已经经过校准或根据一个等量的逐级量度被放置在某个测量连续体上。通俗地说，题库中的题项经过校准后，其难度已具备了等量特征，它们在难度量表上每上升或下降一个单位，就等同于它的相对难度增加或降低了一级。这个“级”不论发生在量表的什么位置，都是等量的。

(4)一个真正意义上的题库允许增加或减少库中的题项数目而不损害原有或剩余题项的作用，或是改变题项分级标记的意义。

(5)受试对题库中用于生成普遍能力测量的题项的回答必须能够累积反映那种能力。也

就是说，题项分数相对于要测量的那种能力来说，必须是一维的，累加的。题项分数越高，所要测量的那种能力也就应该越强。

此外，在严格按照规范建设题库的同时，还应考虑一些能对题库的效度产生影响的因素。这些因素对纸笔考试不构成影响，但对计算机自适应测试来说却无法回避。

首先要考虑的是考生对计算机的焦虑度(computer anxiety)。有研究表明，对计算机的熟练程度越高，以前使用计算机的经验越多，焦虑度就越低，而焦虑度越高，考生的成绩就越受到影响(Cambre and Cook 1985)。因此，计算机自适应测试必须解决如何把考生对计算机的焦虑度降到最低点的问题。

题库的维度也是应该注意的问题。以题项反应理论为基础的计算机自适应测试要求题库中的题项是一维的，也就是说题项所要测量的是一个单一的、主导的潜在能力。尽管它们测量到的可能是多种能力，但其中占主导地位的能力一定是它们所要测量的能力，否则题库的理论效度就出了问题。比如，如果受试在听力试题上的分数不但包含了他的听力水平，而且还包含了其他同样重要的能力，那么这个题库就失去了它的理论效度。

最后，题项是否适合模型也是影响题库效度的一个重要因素。在已知受试的能力和题项难度的情况下，如果受试的回答和预期不一致，那就发生了模型不适的问题。如区分度太高或太低的题项就不适合单参数模型，容易让低水平受试猜对试题正确答案的题项也会被模型认为不适合。在这种情况下，不适合模型的题项就应该从题库中剔除。

4.2 题库的建设

题库的建设一般要经过如下几个阶段。

(1)计划阶段　在这个阶段要确定和描写所要测量的内容。如果是第二语言测试，就要确定所要测试的第二语言的内容范围。由于计算机自适应测试对每一个受试来说内容都是不一样的，题项的选择是根据受试表现出的能力确定的，因此测试内容要覆盖所有的能力层次，以保证计算机为不同能力的受试挑选出能满足内容要求的题项。从这一点不难看出，好的题库所需要的题项数量是很大的。正如 Weiss(1985)指出的，“计算机自适应测试只有在其题库包含了大量的具有良好区分度的题项，并且它们在难度——能力水平线上呈均等分布的情况下才能发挥最佳效果”。

(2)题项编写及试用阶段　题项编好汇总后，首先要经过试用。这时需要题项反应理论来确定题项的好坏，如是否适用，是否需要修改或弃用等。最常用的题项反应理论模型是单参数、双参数和三参数逻辑模型。

(3)校准阶段　在这一阶段，编好并经过试用的题项要拿到受试中试测，然后根据受试对题项的回答测算题项的各项指标，如难度水平、区分度以及猜测指数。这些指标最后都要作为题项参数用于最后试卷的题项选择。需要注意的是，在试测时，受试样本一定要能充分代表目标总体，否则会严重影响自适应测试的效度(Micheline, Chalhoub-Deville and Craig Deville 1999, J. Charles Alderson 1986, Grant Henning 1991)。

五　计算机自适应测试的优势与局限性

由于计算机自适应测试属于计算机化的考试(CBT, computer-based test)，因此，它完全体现了计算机的优点，同时还具备了自适应性测试所独有的优势，归纳起来有如下几点(Chalhoub-Deville and Deville 1999, Henning 1991、1984, Alderson 1986, Tung 1986, Stevenson

and Gross 1991)。

(1)计算机技术允许对个人实施单独测试,减轻了统一考试对时间和监考的压力。

(2)测试实施的条件更加标准化。

(3)受试能即时得到考试结果。

(4)计算机能收集和储存有关受试答题的各种信息,如答题时间,答题策略,是否略过未做某些题项,使用求助信息的次数等。

(5)测试保密程度提高,不用再担心试卷在运输过程中丢失或考生偷题等问题的发生。

(6)使残疾人参考更加方便。例如,对于视力有残疾的考生,计算机可采用放大字体或发声的形式提供试题。

(7)只关注一个考生的能力水平。在传统的纸笔考试中,题项的数量对所有的考生都是一样和固定的,而且要测量的是多种能力。而计算机自适应考试是从一个大题库中选出只和受试能力水平相适应的题项,因而它只需要较少的试题量就能达到传统纸笔测试的测量精度。

(8)不要求考生回答太难或太容易的问题。它选出的每一个题项都符合考生的能力水平。

(9)它的规则系统增加了测试的安全性。由于每个考生所回答的题项都是根据他的能力水平挑选出来的,因而即使是邻座的考生也没有互相抄袭的可能。

(10)它比传统测试更公正、更平等,更有利于提高少数民族和多数民族考生的参考动力,减少多数和少数群体之间测试平均分数的差异。

(11)教师能即时得到学生的成绩,可以更及时容易地发现学生的学习问题。

当然,计算机自适应测试也不是十全十美的,和其他的测试方法一样,它也有自身的缺点。

(1)它的题库要求大量的题项,因此也需要众多的测试工作者对题项进行试用和校准,这个工作量是相当大的。

(2)把纸笔考试的题项移植到计算机上需要进行比较研究,来评估测试介质的变化所带来的潜在影响。另外,还需要对考生进行培训,使他们熟悉机考的程序。

(3)它无法评估扩展性答案,如作文。尽管计算机可以收集受试在这些题项上的表现,但最后的评分还需要人的判断。因而,它通常被局限于评估受试的知识和技能,而不能评估其语言应用能力。

(4)它的开发非常复杂,而且财力上的投入也比较大,要求高水平的心理测量专家和计算机专家参与。

(5)它要求更多的后勤保障。传统的纸笔考试只需要一间大教室就能测试很多人,而计算机自适应测试则需要一个计算机房,以及相匹配的硬件和软件,而且开放的时间必须十分灵活,以满足考生不同的时间要求。

尽管计算机自适应测试还有种种不尽如人意的地方,但从测量的准确性、标准化以及方便程度来讲,它的优势是巨大的,代表了大规模测试的发展方向,因而我们应该及早着手,积极地开展相关方面的研究,尽早地在我国的大学等级考试中引入这种新的测试形式。

六 结语

计算机自适应测试应用于语言测试主要依赖于计算机技术的发展、潜在能力理论的应用以及题库的建设。普及计算机自适应测试的主要障碍是技术和资金。中国是一个外语考试大国,每年仅参加大学英语四、六级统考的在校学生及社会人员就达百万之众。如何利用现代化

的测试手段及测试分析技术为考生提供更方便、更公平、更准确的语言测试是广大测试工作者和组织机构必须要面对的问题。目前我国的大学计算机教育已比较普及，国内的软件开发水平也已达到了相当的高度，更重要的是，我们在大规模外语测试的题库建设、试题编写、数据分析以及测试的组织实施方面都已积累了不少的经验，这些都为计算机自适应测试在我国的实施提供了物质和技术上的准备。如果措施得力，相信在我国开发计算机自适应外语测试系统并投入使用应当不会是太久远的事情。

参考文献

Aderson, J. C. 1986 Computer in Language Testing. In G-Leech (Eds.) *Computers in English Language Teaching and Research*. Longman Group UK Limited.

Cambre, M. A. & Cook, D. L. 1985 Computer Anxiety: Definition, Measurement, and Correlates. *Journal of Educational Computing Research*, 1.

Chalhoub-Deville, M. (Eds.) 1999 *Issues in Computer Adaptive Testing of Reading Proficiency*. New York: CUP.

Chalhoub-Deville, M. 1999 Computer Adaptive Testing in Second Language Contexts. *Annual Review of Applied Linguistics*. 19.

Dunkel. P. (Eds.) 1991 *Computer-Assisted Language Learning and Testing*. New York: Newbury House.

Green, B. F., Jr. 1970 Comments on Tailored Testing. In W. H. Holtzman (Eds.). *Computer-Assisted Instruction, Testing and Guide*. New York: Harper & Row.

Hambleton, R. K. & Swaminathan, H. 1985 *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.

Henning, G. 1984 Advantages of Latent Trait Measurement in Language Testing. *Language Testing*. 1.

Henning, G. 1987 *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge, MS: Newbury House.

Henning, G. 1991 Validating An Item Bank in a Computer-Assisted or Computer-Adaptive Test: Using Item Response Theory for the Process of Validating CATS. In P. Dunkel (Eds.) *Computer-Assisted Language Learning and Testing*. New York: Newbury House.

Lord, F. M. 1980 *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Laurence Erlbaum Associates.

Lord, F. M. & Novick, M. R. 1968 *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Madsen, H. S. 1991 Computer-Adaptive Testing of Listening and Reading Comprehension. In P. Dunkel (Eds.) *Computer-Assisted Language Learning and Testing*. New York: Newbury House.

Stevenson, J. & Gross, S. 1991 Use of a Computerized Adaptive Testing Model for ESOL/Bilingual Entry/Exit Decision Making. In P. Dunkel (Eds.) *Computer-Assisted Language Learning and Testing*. Newbury House.

Tung, P. 1986 Computerized Adaptive Testing: Implications for Language Test Developers. In C. W. Stansfield (Eds.), *Technology and Language Testing*. Washington, DC: TESOL.

Wainer, H. (Eds.). 1990 *Computerized Adaptive Testing: A Primer*. Hillsdale, WJ: L. Erlbaum.

Weiss, D. J. 1982 Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement*. 6.

Weiss, D. J. 1983 *New Horizons in Testing-Latent Trait Test Theory and Computerized Adaptive Testing*. Academic Press, Inc. xiii-8.

Weiss, D. J. 1985 Adaptive Testing by Computer. *Journal of Consulting and Clinical Psychology*. 53.

On the Working Mechanism of CALT

Abstract In this paper several aspects of the Computer-Adapted Language Test (CALT) are examined, including the relationship between the Item Response Theory (IRT) and CALT, the construction of CALT Item Bank. Based on this, the working mechanism of CALT, as well as the advantages and disadvantages of its application is exposed. It concluds that to exploit CALT is general trend.

Key words CALT; IRT; Item Bank

（张宝钧　100083　北京语言大学外语学院英语系）

（责任编辑　王正刚）

书　讯

《诸城方言志》，钱曾怡、曹志耘、罗福腾著，吉林人民出版社 2002 年 12 月出版，25 元。

《宁津方言志》，曹延杰著，中国文史出版社 2003 年 2 月出版，28.60 元。

《舟山方言》，方松熹著，中国文联出版社 2002 年 1 月出版，26 元。

《方言平议》，汪平著，华中科技大学出版社 2003 年 3 月出版，16.80 元。

《沪语盘点》，钱乃荣著，上海文化出版社 2002 年 9 月出版，16 元。

《吴语研究——第二届国际吴方言学术研讨会论文集》，上海市语文学会、香港中国语文学会合编，上海教育出版社 2003 年 1 月出版，50 元。

《海南屯昌闽语语法研究》，钱奠香著，云南大学出版社 2002 年 8 月出版，20 元。