

基于IRT的大学英语词汇在线自适应测试系统的设计

赵传海 吴敏 叶艳

(中国科学技术大学 现代教育技术中心, 安徽合肥 230026)

【摘要】如何科学有效地测量学习者的词汇量, 以及测量其对词汇的掌握程度是当前语言研究者十分关注的问题。文章根据词汇的广度、深度之间的相关性, 提出了在广度测试的基础上进行深度测试的思想, 并将项目反应理论的测试方法、设计思想, 应用到实际测试系统中, 最终设计实现了基于项目反应理论的大学英语四、六级在线自适应单词测试系统。

【关键词】词汇测试; 广度测试; 深度测试; IRT; 单词库

【中图分类号】G434

【文献标识码】B

【论文编号】1009—8097 (2008) 12—0087—04

一 引言

语音、词汇和语法是语言的三大要素。学习语言的最终目的是为了交际, 词汇是语言交际的核心。对于ESL (English As A Second Language) 学习者, 词汇是外语学习的主要瓶颈与最大障碍。在国内, 大学英语考试 (College English Test) 是教育部主管的一项全国性的教学考试, 其中四级考试 (CET-4) 自从1987年, 六级 (CET-6) 自从1989年在我国实行以来, 其目的是在于准确地衡量我国在校大学生的英语综合应用能力, 为实现大学英语课程教学目标发挥积极作用。其对学习者词汇量的要求又是针对大学英语教学大纲而制定, 大致为4500个单词700个词组 (CET-4) 和5500个单词与1200个词组 (CET-6), 词汇量水平以及掌握程度在很大程度上是直接影响四、六级成绩的首要因素。因此, 如何科学有效的测量学习者的词汇量, 以及对词汇的掌握程度的研究成为语言研究者在教学研究活动中十分关注的问题。

为了帮助学习者进行词汇的记忆与学习, 以便有效地进行词汇测试, 作者构建了基于IRT的大学英语词汇在线自适应测试系统, 其中的自适应算法设计是该系统的核心问题。本文首先论述词汇广度和深度的内涵及其测试; 其次引入项目反应理论 (Item Response Theory, 简称为IRT) 以及词汇知识衡量等级 (Vocabulary Knowledge Scale, 简称为VKS) 来进行单词的广度与深度测试; 再次详细介绍了词汇测试系统的设计实现; 最后提出了本系统的一些不足以及今后的研究方向。

二 词汇广度与深度及其测试

词汇测试分为广度测试 (Vocabulary Breadth Measures) 和深度测试 (Assessment of The Depth of Vocabulary Knowledge)。广度测试是估计语言使用者的词汇总量, 深度测试是了解语言使用者对词汇知识掌握的程度。词汇量以及词汇深度知识均可有效预测语言综合能力 (包含听力、阅读、

完型、写作, 下同), 其中词汇深度知识对语言综合能力的预测能力强于词汇广度知识, 特别体现在四、六级的完型填空与写作的预测中, 而总体来说词汇广度与深度呈高度正相关^[1]。

词汇量测试, 一般称为广度测试, 其重要性以及与语言综合能力关系的研究成果颇多, 国外具有代表性的是词汇量与阅读 (Koda 1989; Laufer 1989, 1992; Laufer & Nation 1996; Qian 1999, 2002)^[2-7]及语言综合能力 (Meara & Jones 1988)^[8]呈显著正相关关系。国内具有代表性的有词汇量与语言综合能力成高度正相关关系 (桂诗春 1983, 1985)^[9-10]。当前常用单词量测试的方法有以下几种: 一是概率统计法。即一定样本中随机抽取单词, 选择其正确的意思, 根据其抽样单词答对百分比来做样本总量的推断; 二是词表是否测试法。即认识为是, 不认识为否; 三是Nation (1983, 1990)^[11-12]的分级词汇测试法等。

学习者对词汇知识 (深度) 的习得, 是一个由不同层面和水平组成的连续体, 而不是一个“习得”或“未习得”、“知道”或“不知道”的简单二分的过程。词汇深度有以下的分类方法: 从多个维度 (Dimensional Approach) 界定, 主要代表人物有Cronbach, Richards, Nation, Qian^[13-15]^[6]; 从发展的角度 (Developmental Approach) 出发, 主要代表人物有Dale, PARIBAKHT & Wesche^[16-17]; 主要的测试工具有新西兰维多利亚大学瑞德John Read设计的词汇联想测试 (Word Associate Test) 和PARIBAKHT & Wesche的词汇知识衡量等级 (VKS) 等。

三 项目反应理论 (IRT) 与词汇知识衡量等级 (VKS)

一直以来, 学生为了备考大学英语四、六级考试所做的第一件事往往是背单词。大多数学生仅是单纯的背诵单词的拼写, 对词组的记忆, 而忽略了如何将单词与语法、句法联系起来运用。综合作者所做的文献调研得知, 如何有效科学地进行单词量评估与施测, 以及对一定单词量 (广度) 的基础上再进行“质” (深度) 的测试目前还没有一个行之有效的方法。

上文提及的几种词汇量测试方法，即词汇广度测试，均不能体现学习者的能力特征，从而很难保证学习者的测试效率。其次测试中的施行效率也不高，即抽取的样本量以及如何抽取等。因此本系统采用当前测试中普遍使用的项目反应理论（IRT）^[18]来进行单词量自适应测试。由于国内外单词量测试题型主要有选择题与翻译，而前者应用更广，且具有较高的信度、效度（娄喜祥 2005:2）^[19]，故而本文的单词量测试也采用选择题。然后在此基础上运用 PARIBAKHT & Wesche 的 VKS 工具再进行深度测试。

项目反应理论（IRT）最大优越性在于测试系统可以主动适应受测者状况的“因人施测”问题。试题参数的估计独立于被试样本，而能力参数的估计又独立于试题样本。也就是说，项目反应理论中的这些参数具有不变性，它们不随被试的样本而变化，从而提高了测试效率和测试效度，以及避免了测试过程中被测能力与题目难度的密切关系。理论中最常用的是拉希模型、双参数和三参数逻辑斯蒂（Logistics）模型，运用极大似然法或贝叶斯方法来估计项目的参数难度—区分度和伪随机参数。本文系统采用了三参数逻辑斯蒂模型以及极大似然法，其中三参数逻辑斯蒂模型的函数表达式如下：

$$p_i(\theta) = \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}} \quad (3-1)$$

- 上式中， $p_i(\theta)$ 表示能力水平为 θ 的人答对题目 i 的概率；
 θ ：表示受测者的能力水平；
 D ：表示量表因子， $D=1.702$
 e ：表示自然对数的底， $e=2.71828$
 a_i ：表示题目 i 的区分度；
 b_i ：表示题目 i 的难度；
 c_i ：表示题目 i 的猜测度；

计算机化自适应测试（Computerized Adaptive Testing，简称为CAT）是建构在项目反应理论（IRT）基础上的，从题库的建设、参数的估计到试题的选择再到最后评分，都是以此为指导进行的。由于理论分析和实践经验都证明，只有当题目难度跟受测者水平相适应时，题目所提供的信息量才最大，受测者的积极性最高，误差最小，测试效度才会最高。CAT的核心思想是：系统会根据答题情况不断计算受测者的能力值及信息量，并实时地根据这些参数调整出题策略，选取与受测者能力相对应的试题，最终给受测者的能力与特质一个恰当的评价。

下面介绍下本文采用的单词深度测试算法，即广泛应用的PARIBAKHT & Wesche的VKS工具，该工具使用五个等级将自述与所表现的语言能力结合起来以得出研究对象对各个词的掌握程度，该表包括五项，每项意义如下表1：

表1 (Vocabulary Knowledge Scale)

项目	得分	得分意义
I	1	不认识该词
II	2	见过但不知道意思
III	3	能给出词义
IV	4	用词造句，意义恰当
V	5	用词造句，意义和句法正确

四 词汇测试系统的设计

本系统是在大学英语四、六级单词库的基础上，首先应用IRT理论进行大学英语词汇的广度测试，然后使用IRT工具来进行词汇深度测试，并最终反馈给受测者关于词汇量与词汇掌握程度的度量结果。

其中单词量测试的具体流程如下，先根据受测者的能力初值从单词库中选取一个单词实施测试，如果受测者答对了就选取一个较难的单词再施测；如果受测者答错了就选取一个相对简单的单词再施测；不断重复测验过程，一直到受测者的能力值被精确估计出来为止。

本系统对于受测者有三种方式确定能力初始值。其一，选择历史记录，系统会自动选取该受测者最近的一次测试记录最终能力值作为初始值施测；其二，选择自定义初始值，系统将根据受测者自定义的初始能力值选取单词；其三，选择随机单词测试，这样系统会随机抽取一组单词，根据受测者的答题结果利用极大似然法初步估计其特质水平，然后继续施测。这里要注意的是，如果抽取的一组单词全对或者全错，会再次抽取一组施测，如果仍然是全对或全错，则说明题库中的试题对于受测者而言太难或太易，无法测出其真实水平，此时则终止施测，并向受测者反馈预测结果。在施测过程中，系统会根据受测者的答题结果动态评测其能力值，然后动态调整与之对应的单词难易程度。对于如何终止施测，也有几种方法，如题目数控制、测验估计精度、或者强制退出等。考虑到测量的精度需求以及效率、避免随机猜测等方面，本系统采用了受测者最后3次的估算能力值误差，如果此3次误差值皆小于指定误差范围内，则可以终止施测。其中选择随机单词测试的具体流程图和使用本系统单词量测试生成的能力值曲线图分别参见图1和图2所示：

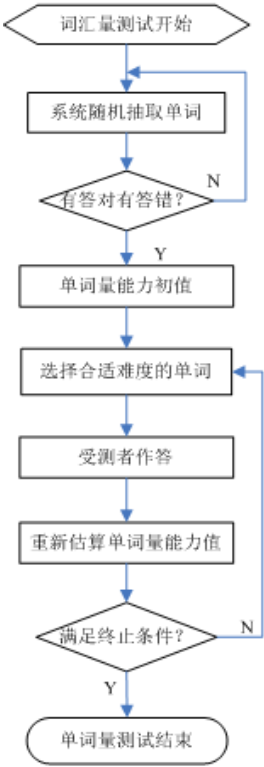


图1 词汇量测试流程图



图2 能力值曲线图

然后，对于能力值的解释，本系统采用了极大似然法估算能力值，其近似正态分布的，从而对能力终值通过一个线性转换，其分布仍是正态，并得到其置信区间值，对应于本

系统即是单词量的范围。最后按照单词的频率高低内选取前1%的单词再进行VHS深度测试，可进一步测试受测者单词的掌握程度，并反馈给受测者。整个测试流程如图3所示：

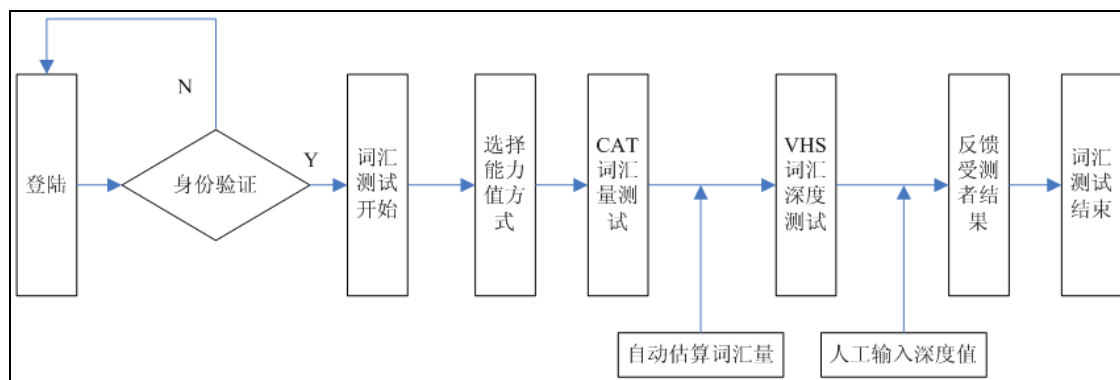


图3 词汇测试流程图

本系统的单词库建立是一个关键，要确定每个单词的难度、区分度、猜测度等参数，而其中难度又尤为重要。通常做法可以根据大纲要求的单词，然后按历年四、六级考试的词频划分，再与专家审核相结合后确定难度值等，或者由样本测试后统计分析确认参数值，本系统综合此两种方法，实现参数动态维护，更好的实现单词库的本身自适应，有效改进结果精度与测试效率。

五 结束语

本文就当前词汇测试提出了一个新的思路与尝试，即在自适应测试单词量的基础上再进行单词的深度测试，让学习者更方便有效地进行自身单词量的评估，进行下一阶段的复习。不仅可以做到“因人施测”，大幅提高测试效率，还可以反馈给受测者对于不同单词的掌握程度，更好的应用于大学英语教学改革。目前系统尚有许多不足，比如没有引入多值法自适应测试、受测者答题时间对于能力值的影响、以及与常用的概率统计等单词量测试方法的比较分析，此外，单词广度与深度对于低级、中级、高级词汇量的学习者所体现的不同相关度的介入等等，这些都值得我们的进一步深入研究与探讨。

参考文献

- [1] 李晓.词汇量、词汇深度知识与语言综合能力关系研究[J].外语教学与研究,2007,39(5):424-450.
- [2] Koda, K.The effects of transferring vocabulary knowledge on the development of L2 reading proficiency [J].Foreign Language Annals,1989,22(4):529-540.
- [3] Laufer, B. A factor of difficulty in vocabulary learning:Deceptive transparency [J].AILA Review,1989,6(1):10-20.
- [4] Laufer, B.How much is necessary for reading comprehension? [A] In H. Bejoint & P.Arnaud(eds.).Vocabulary and Applied Linguistics [C].London:MacMillan,1992:126-132.
- [5] Laufer, B.& P.Nation.Vocabulary size and use:Lexical richness in L2 written production [J].Applied Linguistics,1996,16(3):307-322.
- [6] Qian D D.Assessing the Roles of Depth and Breadth of Vocabulary Knowledge in Reading Comprehension [J].The Canadian Modern Language Review,1999,56(2):282-307.
- [7] Qian D D.Investigating the Relationship between Vocabulary Knowledge and Academic Reading Performance:An

Assessment Perspective[J].Language Learning,2002,(52):513-536.

- [8] Meara,P.& G.Jones.Vocabulary size as a placement indicator [A].In P.Grunwell (ed.).Applied Linguistics in Society [C].London:Center for Information on Language Teaching and Research,1998:80-87.
- [9] 桂诗春(编).中国学生英语词汇量调查,公共外语教学研究文集 [C].上海:上海外语教育出版社,1983.
- [10] 桂诗春.我国英语专业学生词汇量的调查与分析[J].现代外语,1985,(1):1-6.
- [11] Nation I S P.Testing and teaching vocabulary [J].Guideline,1983,5(1):12-25.
- [12] Nation I S P.Teaching and Learning Vocabulary [M].Victoria University of Wellington:English Language Institute,1990.
- [13] Cronbach L J.An Analysis of Techniques for Diagnostic Vocabulary Testing [J].Journal of Educational Research,1942,(36):206-217.
- [14] Richards J.The role of vocabulary [J].TESOL Quarterly,1976,(10):77-89.
- [15] Nation I S P.Learning Vocabulary in another language [M].Cambridge,England:Cambridge University Press,2001.
- [16] Dale E.Vocabulary Measurement:Techniques and Major Findings [J].Elementary English,1965,(42):895-907.
- [17] PARIBAKHT T S,W ESCHE M B.The relationship between reading comprehension and second language development in a comprehension-based ESL program [J].TESL Canada Journal,1993,(11):9-29.
- [18] Howard Wainer & Robert J. Mislevy.Item Response Theory Item Calibration, and Proficiency Estimation.Computerized Adaptive Testing:A Primer Second Edition,2000,4.
- [19] 娄喜祥.两种常用的外语词汇量测试方式的信度及效度对比[J].外语与翻译,2005,(2):220-241.
- [20] 刘绍龙.论二语词汇深度习得及发展特征[J].外语教学与研究,2001,(6):436-441.
- [21] Lawrence M. Rudner. An On-line Interactive Computer Adaptive Testing Tutorial[Z].< <http://EdRes.org/Scripts/cat>, 1998,11.>

The Design of IRT-Based Online Adaptive Testing System for College English Vocabulary

ZHAO Chuan-hai WU Min YE Yan

(University of Science and Technology of China, Center of Modern Educational Technology, Hefei, 230026,China)

Abstract: How to measure learners' breadth of vocabulary scientifically and the depth in mastering vocabulary is a major issue which many language researchers are focusing on currently. This article introduces the correlation between the breadth and depth of vocabulary, and presents the idea of measuring the depth of vocabulary based on the measurement of breadth. Finally the online adaptive vocabulary testing system is achieved for College English Band-4 and Band-6, by applying the testing method of Item Response Theory (IRT) and design thoughts to practical testing system.

Keywords: Vocabulary Testing; Breadth Measuring; Depth Measuring; IRT; Vocabulary Bank

(上接第 110 页)

Research and Practice on the Model of Two Levels Physics Virtual Experiment in Middle School

HU Shan¹ YANG Chun²

(1.Educational Technology Service, SiChuan International Studies University,Chongqing,40031,China; 2.Software Key Lab of Sichuan, Sichuan Normal University, Chengdu,610066,China)

Abstract: There are several disadvantages in traditional middle school physics experiment educational model.For improving the traditional model, a solve method is proposed from the view point of information technology and education, which is designing and developing the real three-dimensional physics virtual experiment. After analyzing the combination of middle school physics experiment teaching aims and virtual technology, a classification method for middle school physics virtual experiment is proposed in the first, then on this base and with the combination of "Learning condition theory", the two level middle school physics virtual experiment model is proposed.

Keywords: Three-Dimension; Middle School Physics; Virtual Experiment; Model