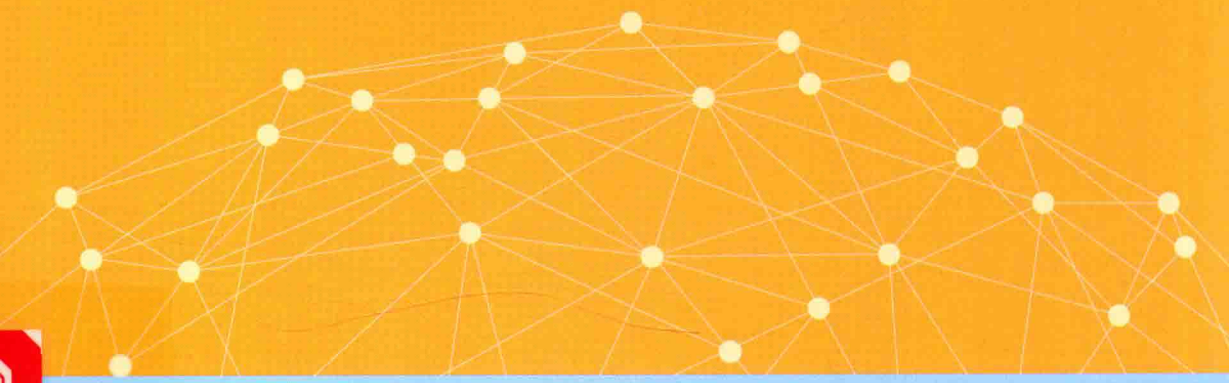
计算机自适应英语语言能力测试系统设计与效度验证

以TEM 4词汇与语法题为例

A DESIGN AND VALIDATION OF COMPUTERIZED ADAPTIVE ENGLISH LANGUAGE PROFICIENCY TEST SYSTEM

A STUDY ON VOCABULARY AND GRAMMAR QUESTIONS OF TEM 4

张一鑫 编著



中国纺织出版社有限公司

国家一级出版社
全国百佳图书出版单位

计算机自适应英语语言能力测试系统设计与效度验证

以 TEM 4 词汇与语法题为例

A DESIGN AND VALIDATION OF COMPUTERIZED ADAPTIVE ENGLISH LANGUAGE PROFICIENCY TEST SYSTEM

A STUDY ON VOCABULARY AND GRAMMAR QUESTIONS OF TEM 4

张一鑫 著

中国纺织出版社有限公司 | 国家一级出版社
全国百佳图书出版单位

图书在版编目（CIP）数据

计算机自适应英语语言能力测试系统设计与效度验证：以TEM4词汇与语法题为例 / 张一鑫著. -- 北京 : 中国纺织出版社有限公司，2019.11

ISBN 978-7-5180-6510-3

Ⅰ. ①计… Ⅱ. ①张… Ⅲ. ①英语－计算机辅助测试－自适应程序－研究 Ⅳ. ① H319-39

中国版本图书馆CIP数据核字（2019）第168177号

责任编辑：郭 婷 责任校对：王花妮 责任印制：储志伟

中国纺织出版社有限公司出版发行

地址：北京市朝阳区百子湾东里A407号楼 邮政编码：100124

销售电话：010—67004422 传真：010—87155801

http://www.c-textilep.com

E-mail:faxing@c-textilep.com

中国纺织出版社天猫旗舰店

官方微博 http://www.weibo.com/2119887771

三河市宏盛印务有限公司印刷 各地新华书店经销

2019年11月第1版第1次印刷

开本：710×1000 1/16 印张：12.25

字数：250千字 定价：56.00元

前 言

随着测量理论和计算机技术的不断发展，计算机自适应语言测试系统（computerized adaptive language test system，简称 CALT）的开发与应用是近年来国内外语言测试领域研究的核心热点方向。相对于传统的纸笔语言测试（paper-and-pencil language test，简称PPLT）或普通的计算机辅助语言测试（computer-based language test，简称 CBLT），CALT 有更加人性化、更高的精确性、出题更加灵活、施考及评分更简洁以及测试效率更高等优势。

本研究的主要目的是：1）完成计算机自适应测试大型题库的构建，该题库包括以下三种题型：完形填空题、语法选择题和词汇选择题。在内容方面，该题库涵盖英语专业四级考试（TEM 4）要求的所有语法知识点和考试大纲所要求的所有词汇。2）在题库的基础上设计一个计算机自适应英语测试系统，并采用“评估使用论证”（assessment use argument，简称 AUA）对该测试进行效度验证。

针对第一个研究目的，采用的主要研究方法是对 1997 年至 2007 年的英语专业四级考生数据进行数据处理和分析，筛选出符合题库建设需求的题目，编写入题库。使用 AUA 框架对本过程进行效度验证，并回答本阶段研究所讨论的四个有关研究中题目是否适合 IRT 理论的问题。采用以下分析方法对数据进行处理：1）使用 SPSS20.0 进行探索性因子分析，并用 AMOS7.0 进行验证性因子分析，以检验单维性假设；2）用 Bilog2.0 对项目局部独立性假设进行检验；3）使用 Bilog2.0 中的 2PLM 对二元计分项目进行分析，用 GRM 及 GPCM 对多元计分项目进行分析；4）使用 Bilog2.0 以及 SIBTEST 对项目进行性别 DIF 检验，然后对存在性别 DIF 的项目进行内容分析，以确定该项目是否需要从题库中剔除。

针对第二个研究目的，设计三个步骤：CALT 设计、模拟、运行与效度验证。在 CALT 设计方面 :1）采用以下方法：测试按照完形填空题、语法选择题和词汇选择题的顺序进行；2）项目选择采用最大信息量选择法（maximum information，

简称 MI），并兼用内容平衡、曝光控制等措施；3）能力估计采用贝氏期望后验法（Expected A *Posteriori*，简称 EAP）；4）终止原则同时采用标准误差控制原则和总题量控制原则。在模拟方面采用了 Firestar 与 R 软件进行四次模拟运行。针对 CALT 运行与效度验证，采用的数据分析方法包括：1）SPSS20.0 进行配对 T 检验，用 AMOS7.0 进行验证性因子分析；2）使用 AMOS7.0 对计算机熟悉度、TEM 4 成绩、CALT 成绩进行结构方程模型建模；3）使用 AMOS7.0 对以上模型进行多群组结构方程模型分析。

本研究不仅设计一个涉及语法与词汇测试的 CALT，填补了国内语言测试领域的研究空白，而且在国内外语言测试领域中首次尝试对 CALT 进行了较为系统的效度验证。在计算机技术已经日趋完善并对语言测试带来变革的今天，本研究在理论和实践上均具有重要意义。

在理论方面，本研究的效度验证不再局限于 CALT 与 PPLT 之间的等效性，而是倡导更好地理解 CALT 所考查的构念，理解计算机熟悉度、语言能力与 CALT 所考查的构念之间的关系。其次，本研究尝试在测试分数解释方面使用 AUA 框架，在一定程度上促进基于论据的效度验证方法在语言测试领域得到更广泛的应用。

在实践方面，本研究介绍了 CALT 开发过程中的详细步骤，指出了以往研究在 CALT 题库建设方面存在的不足，有利于扩展 CALT 开发方面的知识，并能给考虑开发 CALT 系统的机构提供实证依据。其次，本研究也探讨了计算机熟悉度（computer familiarity）对考生在 CALT 中表现的影响，这将有助于 CALT 开发者与使用者更积极主动地预测 CALT 实施过程中可能遇到的问题，针对性地采取解决方案，并更适当使用 CALT 的测试结果。

不可否认，本研究存在一定的局限性，未来研究可以从以下几个方面着手：1）本研究采用了多元计分 IRT 模型解决了局部独立性假设违反的问题，但是此方法的一大缺陷是项目层次信息的丢失，未来研究可以考虑采用多维 IRT 模型，如双因素模型（bi-factor model）、题组反应理论模型（testlet response theory model），进行项目参数估计，从而保证项目层次信息的完整性；2）本研究采用

的单独分析使同时参与分析的项目数量减少，在一定程度上加大了参数估计的误差，未来研究可以考虑彩双阶全信息项目因子分析模型进行参数估计，真正实现多维 CALT 创建；3）本研究仅从性别角度考查了 CALT 因子结构的一致性，未来研究可以从考生专业、生源地等角度着手；4）本研究主要基于定量分析研究 CALT 效度，未来研究可以采用定性分析方法研究考生在 CALT 中的认知过程，从而更好地探讨 CALT 的效度问题；5）本研究设计的 CALT 仅包括语法和词汇部分，而未涉及听力、阅读、写作与口语考试，随着自动评分技术的发展，未来研究可以尝试设计更全面的考查语言能力的 CALT。

Perface

With the fast development of the measurement theory and computer technology, the development and application of computer adaptive language test system(CALT) has been the hot spot of language testing of both domestic and abroad research field. Compaied with traditional paper-and-pencil language test(PPLT) and ordinary computer-based language test(CBLT), CALT has many advantages, such as more user-friendly, more accurate, more flexible in ways of presenting test items, more convenient in ways of administrating test and scoring, and more efficient.

The purposes of this study were to: 1) design a computerized adaptive language test (CALT) to assess grammar and vocabulary proficiency in English using mixed-format with dichotomous and polytomous item response theory (IRT) models, and 2) to investigate the validity of the CALT under the assessment use argument (AUA) framework.

In the process of item bank construction, data of all English majors in China who took part in the TEM 4 from 1997 to 2007 were thoroughly analyzed with Bilog 2.0, AMOS 7.0 and SPSS 20.0 software. The responses were used for item calibration and differential item functioning (DIF) detection. Research methods include: 1) exploratory factor analysis (EFA) with SPSS 20.0, confirmatory factor analysis (CFA) with AMOS 7.0, to examine the uni-dimensionality assumption; 2) examine the local dependent assumption with Bilog 2.0; 3) use 2PLM to examine dichotomous items and use GRM and GPCM to examine polytomous items with Bilog 2.0; 4) DIF test with Bilog 2.0 and SIBTEST.

For the second research objective, there are three procedures: CALT design, simulation, processing and validation. In the design of CALT, arrange the sequence of cloze, grammar choice and vocabulary choice; use maximum information(MI) method to select items, and consider content balance and exposure control; use expected a posteriori(EAP) to estimate ability; use a combination of variable-length stopping rule and fixed-length stopping rule. In the simulation of CALT, the Firestar and R software are employed. In the processing and validation of CALT, analyses include: 1) T test of SPSS

20.0 and CFA of AMOS 7.0; 2) SEM of AMOS 7.0 to examine the computer familiarity, TEM 4 results and CALT results; 3) analyses of models above with AMOS 7.0.

This study has investigated the procedures used to develop a CALT designed to assess grammar and vocabulary proficiency in English with mixed-format, and examined the validity issues of the CALT within Bachman and Palmer's (2010) AUA framework. Major findings of the study are summarized as follows, in order of the three stages, namely, item pool construction, overall CALT design, and CALT validation.

Theoretically, the present study, for the first time in the literature, fully investigates the construct validity of a CALT. Construct validity of CALTs has not been fully investigated before, possibly due to the fact that no consensus has been reached as to what the CALTs measure.

Practically, the present study provides insight into the specific procedures that need to be followed in the development of CALTs and points out a few key issues, such as DIF detection, that were ignored in previous CALT development research. The trend of using computerized language tests in large-scale language assessment in the world, combined with the power of CALTs in discriminating test takers in an effective way, makes this a critical area of study from a practical standpoint.

There are limitations in this study, which I hope could inspire other researchers in future study. For instance, the possibility of applying one of the multi-dimensional IRT models in item calibration could be further explored; more flexible tow-tier full-information item factor analysis model to calibrate the grammar and vocabulary sections simultaneously could be applied; future studies should also attempt to incorporate other variables as mediating factors of the influence of computer familiarity on test takers' performance in the CALT, etc.

List of Symbols and Abbreviations

AIC	akaike information criterion
AUA	assessment use argument
BIC	bayesian information criterion
CACC	computer access
CAIC	consistent version of this statistic
CALT	computerized adaptive language test/testing
	computer-adaptive language test/testing
CALTEA	english ability measured in CALT
CAT	computerized adaptive test/testing
CATT	computer attitude
CBLT	computer-based language test/testing
CCC	category characteristic curve
CET	college English Test
CETGEP	general English proficiency measured by CET4
CEXP	computer experience
CF	computer familiarity
CFA	confirmatory factor analysis/analyses
CFI	comparative fit index
CTT	classical true score theory
DBF	differential bundle functioning
EAP	expected a *posteriori*
EFA	exploratory factor analysis/analyses
ESL	english as a second language

F_nGC_m	themth grammar choice item in the nth form
F_nVC_m	themth vocabulary choice item in the nth form
F_nCL_m	themth cloze item in the nth form
GFI	goodness of fit index
GRM	graded response model
GPCM	generalized partial credit model
IB-CET	internet-based College English Test
ICC	item characteristic curve
IELTS	international English Language Testing System
IRT	item response theory
IRT-LRT	iRT likelihood ratio test
KMO	kaiser-Meyer-Olkin
GC	grammar choice item
VC	vocabulary choice item
MCQ	multiple-choice questions
MI	maximum information
MLE	maximum likelihood estimation
PCM	partial credit model
PPLT	paper-and-pencil language test/testing
RMSEA	root mean square error of approximation
SE	standard error
SIBTEST	simultaneous item bias test
TOEFL	test of English as a Foreign Language
1PLM	1-parameter logistic model
2PLM	2-parameter logistic model
3PLM	3-parameter logistic model

目录

Chapter 1 Introduction

CALT is the most sought direction in the development of modern assessment and the research emphasis in educational-purposed measurement. This chapter demonstrates the background of this study starting from the paper-and-pencil based language test (PPLT) to the computer-adaptive language test (CALT), and then formats the basic research question, research orientation, and research content. The research methodology and research implications are also briefed in this chapter. At last, there is the organization of this study.

1.1 Research Background

During our past school years, we have all experienced countless tests, through which schools examine the teaching quality and learning results to know the extent of students' mastering of knowledge, consolidate students' learning, and then improve their teaching quality. Therefore, testing is an indispensable component of the school teaching. The traditional PPLT is the most familiar and long-term form, in which teachers should develop test paper, print test paper, send out the paper, collect the paper, score the paper and give feedback to the paper, which, apparently, is a long and complicated process and not suitable for the modern instruction. In addition, due to the same test paper and test cheating, the traditional PPLT has been questioned to some extent of their objectiveness. Therefore, there are several necessary questions, such as what test forms should be adopted and how to improve the quality and efficiency, needed to be concerned by all the educationists.

With the development of test theory and technology, the validity and reliability have been greatly improved. There is CBLT and CALT implemented during the process of teaching and testing. After the implementation of computer science into

the education and testing, many CBLT systems have come into use, such as diploma test, assessment test, certificate test. On the one hand, however, computer is used just as a medium to display, transfer, and store the test content and answers in such CBLT system, which doesn't really exert computer's full function in information management; on the other hand, in the construct of the test, test items are exactly the same but with different sequences, and all these items, to every student, are only useful for some of them, which can truly test students' ability while others are either too difficult or too easy, just like the traditional PPLT. For those with lower ability, the difficult items can barely be answered, which could definitely affect their mood in doing the test and hence the result of their true ability; meanwhile, for those with the higher ability, the easy items can almost be useless, which just occupy the content and waste a considerable amount of time of the test takers. Either way, such CBLT fails to assess the true ability of the test takers to some extent. Therefore, there is another problem for the researchers to wonder, which is how to efficiently and accurately test the true ability of the test takers.

Initiated in the 1980s, CALT, which was once termed as "the new world in testing", could efficiently and accurately assess the true ability of test takers. The assessment of test takers' ability doesn't rely on the items, and the parameter of the test items doesn't rely on the test takers in the CALT. In other words, despite using totally different test items, CALT can also assess the ability of test takers, compare the results of test takers, and finally define the discrepancy of the test takers.

The computer is not just a medium in CALT but a "decision maker" in that it should actively determine to select the corresponding items with good test variety and functionality by assessing the ability of test takers in the test. Put it another way, when assessing test takers with high ability, it selects items with higher difficulty; when assessing those with low ability, it displays items with lower difficulty. The CALT system is not a fixed test but an active and adaptive test, which can adjust the difficulty and test content in accordance with the ability of test takers and assess the true ability of the test takers by using the least number of test items.

Therefore, CALT is now deemed as the most efficient and accurate form of test

and becoming increasingly popular for four major reasons:

1) CALT could save as much as 50% of testing time; 2) with different test items, CALT could greatly reduce the odds of test cheating; 3)CALT has a higher accuracy of ability estimation;4) comparing the test takers' ability in the same scale in CALT would enhance the validity and reliability.

Previous empirical studies concerning CALT in the field of language assessment have mainly focused on its design and validation. In terms of design, the majority of CALTs have been developed to assess grammar and vocabulary, only with a handful being devoted to the assessment of grammar, probably due to the complexity involved in adding a speech component to the CALT development process. In addition, most CALT design studies have mainly focused on the introduction of item pool development, which is the prerequisite of the adaptive algorithms in CALT (Flaughter, 2000). However, the majority of CALT design studies are still limited in terms of item bank development in the following four aspects.

The first part of this study, therefore, is to develop an item bank for CALTs are stand-alone items where one passage is followed by one item (e.g., Brown & Iwashita, 1996; Madsen, 1991; Stevenson & Gross, 1991; Young, et al., 1996; Linacre, 1999; Luecht, 1999; Sumbling, 2007; Nogami, 2010), though the most widely used testing format in real language tests is testlets where a set of items are based on the same passage, especially in grammar and vocabulary tests.

Second, most of the item banks consist of items measuring different aspects of language proficiency such as reading, vocabulary and grammar, but little attention has been given to how different components in the item pool might have influenced the uni-dimensional construct measured by the test.

Third, there is too much reliance placed on Rasch modeling (e.g., Stevenson & Gross, 1991; Brown & Iwashita, 1996; Young, et al., 1996; Dunkel, 1999; Sumbling, 2007), with very few attempts made to explore the possibility of selecting the best model based on model-data fit to conduct item parameter estimation.

Fourth, none of the previous CALT researches has addressed the issue of DIF in item pool construction, even though the occurrence of DIF poses a great threat to the