Foreign Language Culture Teaching

外语·文化·教学论丛

Design and Validation of a Computerized Adaptive EFL Test

计算机自适应语言测试模型设计与效度验证

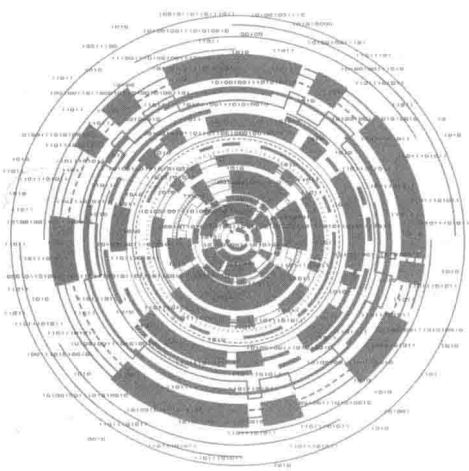
何莲珍　闵尚超　著

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

Design and Validation
of a Computerized Adaptive EFL Test

计算机自适应语言测试模型设计与效度验证

何莲珍　闵尚超　著



ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目(CIP)数据

计算机自适应语言测试模型设计与效度验证 / 何莲珍
闵尚超著. —杭州：浙江大学出版社，2016.12
ISBN 978-7-308-16355-2

Ⅰ. ①计… Ⅱ. ①何… ②闵 Ⅲ. ①电子计算机—自适
应程序—研究—英文 Ⅳ. ①TP31

中国版本图书馆 CIP 数据核字(2016)第 257961 号

计算机自适应语言测试模型设计与效度验证
何莲珍 闵尚超 著

责任编辑 诸葛勤
责任校对 潘晶晶 沈炜玲
封面设计 周 灵
出版发行 浙江大学出版社
(杭州市天目山路 148 号 邮政编码 310007)
(网址：http://www.zjupress.com)
排 版 浙江时代出版服务有限公司
印 刷 杭州杭新印务有限公司
开 本 710 mm × 1000 mm 1/16
印 张 12.5
字 数 224 千
版 印 次 2016 年 12 月第 1 版 2016 年 12 月第 1 次印刷
书 号 ISBN 978-7-308-16355-2
定 价 35.00 元

序 言

桂诗春

由浙江大学何莲珍教授主持完成的国家社科基金项目“计算机自适应语言测试模型设计”优秀结题成果以专著的形式问世，是一件令人兴奋和鼓舞的事。该项目不仅是建立了一个模型，而且还做了许多效度验证，其结果是可信的，为我国的外语考试改革探索提供了一条思路，很值得进一步探讨和完善。

众所周知，我国的考试制度经历了1300多年的实践，科举制度经历过盛衰，然后进入民国时期和中华人民共和国成立后的时期，近年来又出现了对考试的社会性（亦称批评性）思考。考试面临的许多问题都跟经济和文化教育发展不平衡有关，而不是考试本身造成的问题；不从根本上解决这些问题，任何措施都是苍白无力的。考试仅是一种测量工具，其责任就是公平地、准确地测量出被测量者的学识、能力和水平。它们无法缓解经济发展不平衡问题。把社会发展中一些未能解决的矛盾都归咎于考试，是一个错误命题。就考试本身而言，试题的保密也是一个始终困扰着大家的问题。在现代技术支持下的自适应考试所要解决的，就是每一个考生所做的试题在信度、难度和区分度等方面都是一致的，但同时又是来自题库里的不同题目，而且是具有自适应性的。这就是根据其自身水平编制而成，而且是在电脑上完成的。

何莲珍教授所主持的项目包括听力和阅读的题库，而且做了效度检验，验证了：1）计算机自适应语言测试与其他模式语言测试的等效性；2）计算机熟悉度对考生在计算机自适应语言测试中表现的影响；3）计算机自适应语言测试构念在男女考生群体中的一致性。这应该说是首次尝试对计算机自适

应语言测试进行较为系统的效度验证，而且尝试在测试分数解释方面使用“评估使用论据”框架，有利于促进基于论据的效度验证方法在语言测试领域的广泛应用，同时提供了一种研究范式，有利于在未来的研究中更好地探讨基于计算机自适应语言测试分数所作推论的公平性。这些研究对当前一些“假、大、空”的研究具有示范意义。

就研究本身而言，也有一些可以继续努力的地方：1）继续了解“听”和“读”之间的关系，它们都属于接受性能力。2）对产出性能力如“说”和“写”，也也可以做适应性测试的尝试。它们对建库来说，难度不算很大，无非是储存一些题目，但是对评估而言却有很多困难。是人工评估还是计算机评估？机器改作文，目前还不很成熟；机器评估口语，还牵涉到转写或是直接评估的问题。3）目前使用的是项目反应理论的双参数模型，可以尝试同时使用单参数模型与之比较——前者多了一个区分度指数，但后者更节约时间。

目 录

表目录

图目录

第一章 绪论

1.1 研究背景

随着计算机技术与测量理论的不断发展，建立大型的语言测试试题库并基于题库实现计算机自适应语言测试（computerized adaptive language testing）是近年来国外语言测试研究的热点问题。计算机自适应测试兴起于20世纪80年代中期，但直到80年代后期才真正被应用到语言测试领域（Canale, 1986; Henning, 1987, 1991; Meunier, 1994; Chalhoub-Deville & Deville, 1999; Alderson, 2000; Chalhoub-Deville, 2001; Chapelle & Douglas, 2006; Ockey, 2009）。相对于传统的纸笔语言测试（paper-and-pencil language testing）或普通的计算机辅助语言测试（computer-based language testing），计算机自适应语言测试有以下优势：1）测试信度与测试效率高；2）即时反馈效果良好；3）施考安全性好；4）测试的个性化程度高，等等。

计算机自适应语言测试的主要理论依据为项目反应理论（item response theory）。项目反应理论是一组用于阐述考生答题行为与潜在能力之间关系的数学模型，其最大优点是项目数据与样本数据之间具有独立性，即项目参数估计不受其所施测的样本影响，样本能力估计不受其所施测的项目影响。因此，即使考生在测试过程中所得到的考题不一样，仍可以对考生能力进行估计并直接比较，这一优点极大地推动了计算机自适应语言测试的设计与应用。

依据计分模式，项目反应理论可以分为二级计分项目反应理论模型和多级计分项目反应理论模型。二级计分项目反应理论模型中，考生在题目上的得分只有0分、1分两种可能性，二级计分项目反应理论模型包括单参数模型（one-parameter logistic model）、双参数模型（two-parameter logistic model）、三参数模型（three-parameter logistic model）。多级计分项目反应理论模型中，

考生在题目上的得分有0分、1分、2分等多种可能性，常见的多级计分项目反应理论模型有等级反应模型（graded response model）、分部评分模型（partial credit model）和广义分部评分模型（generalized partial credit model）。

项目反应理论的基本假设为单维性与局部独立性，单维性指同一份考卷中的所有题目测量同一种能力。尽管长期以来，语言测试领域在语言能力的单维性问题上争论不休，但目前较为公认的一种观点是单维性是一个度的问题，而非存在与否的问题。局部独立性指考生在各道题目上的答对概率相互独立，即考生的潜在能力是影响考生答题的唯一因素，当排除这个因素的影响后，考生在不同题目上的答题行为之间不存在任何关系。但是在大规模英语测试中，局部独立性假设往往会被违反，因为常见的题型是若干道选择题基于同一篇章。在局部独立性假设被违反的情况下，采用标准的二级计分项目反应理论模型进行项目分析，不仅会导致模型与数据的不拟合，而且会因为对项目区分度的估值过高而导致对测试信息量——即对测量精确度——的过高估值。解决上述问题的一个有效方法是采用多级计分项目反应理论模型。该方法把基于同一篇章的若干题目看成一个整体，即把考生在同一篇章所有题目上的得分相加，作为一个多级计分题目，运用多级计分项目反应理论模型进行参数估计。除项目反应理论以外，计算机自适应语言测试的成功与否主要取决于其四个重要组成部分的功能，即题库、项目选择、能力估计、终止原则。

目前，国内关于计算机自适应语言测试方面的研究基本停留在文献综述或简要介绍上，只有极少数学者进行了计算机自适应语言测试模型设计的实证研究。国外关于计算机自适应语言测试方面的实证研究相对较多，主要探讨计算机自适应语言测试的模型设计与效度验证。下面我们将简要地介绍计算机自适应语言测试模型设计与计算机自适应语言测试效度验证方面的相关研究。

1.1.1 计算机自适应语言测试模型设计相关研究

Larson（1987）是语言测试领域中首例尝试计算机自适应语言测试模型设计的实证研究，在此项研究的基础上，近20多年来研究者们纷纷尝试，并取得了一系列的研究成果。表1.1列出了这20多年来计算机自适应语言测试模型设计方面的实证研究。

表 1.1 计算机自适应语言测试模型设计相关研究

第一作者（年份）	所测技能	所测语言	题库规模	题型	项目反应理论模型
Stevenson (1991)	阅读 语法	英语	170 题	选择题	Rasch
Madsen (1991)	听力 阅读	英语	750 题	选择题	Rasch
Kaya-carton (1991)	阅读	法语	600 题	选择题 填空题	/
Brown (1996)	语法	日语	225 题	选择题	Rasch
Young (1996)	阅读	英语	85 题	选择题	Rasch
Larson (1999)	阅读	荷兰语	/	选择题	/
Dunkel (1999)	听力	豪萨语	144 题	/	Rasch
Laurier (1999)	词汇 语法 阅读 听力 语言运用	法语	/	选择题 填空题	3PLM GRM
Linacre (1999)	阅读	英语	/	选择题	Rasch
Luecht (1999)	阅读	英语	400 题	选择题	Rasch
何莲珍 (2004)	阅读 词汇 语法	英语	431 题	选择题	2PLM
Nielsen (2004)	阅读 词汇 语法	英语	/	选择题 填空题	/
Giouroglou (2005)	语法 词汇 阅读	英语	600 题	选择题 简答题	/
Sumbling (2007)	词汇 语法 听力	英语	775 题	选择题 完形填空	Rasch
Papadima-Sophocleous (2008)	词汇 语法 阅读	英语	1084 题	选择题	CTT
Nogami (2010)	词汇 听力	英语	4000 题	选择题 听写	2PLM 3PLM

注释：2PLM 为双参数模型；3PLM 为三参数模型；GRM 为等级反应模型；CTT 为经典真分数理论

总体而言，在设计方面，绝大多数计算机自适应语言测试采用选择题或填空题的形式考查考生的词汇、语法以及阅读能力，仅有少数涉及听力能力，因为听力测试中的语音成分使得计算机自适应语言测试的模型设计过程更为复杂。大部分实证研究主要介绍题库建设（如 Dunkel, 1999; Sumbling et al., 2007）或设计过程中的决策制定（如 Laurier, 1999），有助于我们更好地了解心理测量模型与计算机技术在语言能力评估中的应用。题库建设之所以成为以往研究关注的核心问题，是因为题库质量事关计算机自适应语言测试所倡导的高测量精度与效度是否能在实际运行中得以实现。没有一个高质量的题库，无论项目选择程序、能力估计方法及终止原则有多科学，计算机自适应语言测试的成功都无法得到保证。尽管如此，过往的研究在计算机自适应语言测试题库建设方面仍存在以下四个方面的局限性：

第一，大多数题库仅采用独立项目，即每道题目基于一个独立的篇章，并采用二级计分项目反应理论模型进行项目参数估计（如 Stevenson & Gross, 1991; Madsen, 1991; Brown & Iwashita, 1996; Young et al., 1996; Dunkel, 1999; Linacre, 1999; Luecht, 1999; Sumbling et al., 2007; Nogami & Hayashi, 2010），而在实际的语言测试中，尤其是听力与阅读测试中，使用最为广泛的题型是题组（testlet），即若干道题目基于同一篇章。题组在之前的计算机自适应语言测试中使用较少的原因是在若干个项目基于同一篇章的情况下，项目之间可能会相互关联，从而导致项目反应理论的基本假设——局部独立性假设——被违反（Rosenbaum, 1988; Sireci et al., 1991）。解决该问题的一个方法是将基于同一篇章的多个项目视为一个多级计分项目，并采用多级计分项目反应理论模型进行参数估计（Rosenbaum, 1988; Thissen et al., 1989; Lee, 1998）。Laurier（1999）的计算机自适应语言测试设计采用了这种方法，为我们提供了一个典型的例证。在该项研究中，Laurier（1999）不仅采用二级计分项目反应理论模型分析独立项目，同时采用多级项目反应理论模型中使用非常广泛的等级反应模型分析题组项目，有效地处理了局部独立性假设被违反的问题。通过采用包含二级计分独立项目和多级计分题组项目的“混合式测试设计”（Lau & Wang, 1998; Rosa et al., 2001），Laurier（1999）的计算机自适应语言测试模型最大限度地模拟了真实测试场景，为计算机自适应语言测试在语言测试领域的应用做出了杰出贡献。其他也有一些研究（如 Young et al., 1996; 何莲珍, 2004）在计算机自适应语言测试题库建设中采用了题组项目，但是这些研究并没有提供局部独立性假设方面的诊断信息，而是将基于同一题组的项目和不基于同一题组的项目混在一起，采用二级计分项目反应

理论模型进行分析，忽略了局部独立性假设可能被违反这一问题，所以其测量准确度仍有待商榷。

第二，虽然大部分题库包括词汇测试、语法测试、阅读测试等多个组成部分，但是很少有研究关注题库中的不同组成部分在多大程度上影响整个题库的单维性。目前语言测试领域的一个共识是：语言能力是多维的，不仅包含一个总的高阶能力因子，还包含若干不同的二阶能力因子（如 Bachman & Palmer, 1981, 1982; Carroll, 1983; Bachman et al., 1990, 1995; Sasaki, 1996; Shin, 2005; Song, 2008）。因此，过往的研究中把不同测试部分（如听力、阅读、词汇与语法）的项目放在一起进行项目估计的习惯做法存在一定的问题。更具体地说，这种做法忽视了不同测试部分可能存在不同的能力因子，模糊了不同潜在能力因子间的界限，从而可能导致项目参数估计和考生能力估计的不准确。因此，更理想的方法是先检查不同测试部分的项目是否与同一测试部分的项目一样符合单维性假设，然后再确定各测试部分项目的校准方式，即合在一起校准或分别进行校准。

第三，在模型选择方面，过往的题库建设过分依赖 Rasch 模型（如 Madsen, 1991; Stevenson & Gross, 1991; Brown & Iwashita, 1996; Young et al., 1996; Dunkel, 1999; Linacre, 1999; Luecht, 1999; Sumbling et al., 2007），极少有研究通过观察模型数据拟合度从一系列理论上可行的模型中选择最佳模型对项目进行参数估计。过往的研究倾向于使用 Rasch 模型的原因各异，有的是因为样本量有限，有的是为了使得计算机自适应语言测试系统运行更为简单。无论基于何种原因，Rasch 模型在计算机自适应语言测试中的过度应用都值得商榷。Rasch 模型的哲学理念与其他二级计分项目反应理论模型（如双参数模型、三参数模型）存在原则上的差别。Rasch 模型以模型为驱动，关注的核心问题是数据是否符合选定模型，如不符合，则说明数据有问题；而其他二级计分项目反应理论模型则以数据为驱动，关注的是所选定的模型是否符合现实数据，如不符合，则更换模型重新拟合（Zumbo & Macmillan, 1999）。换言之，在 Rasch 模型倡导者的眼里，没有不好的模型，只有不好的数据。但是，这个观点在语言测试领域并不适用。我们不能因为某个项目或某个考生的反应方式不符合 Rasch 模型，就放弃该项目或该考生。相反，我们应该从一系列理论上可行的模型中根据模型数据拟合程度选择最佳模型对项目进行参数估计，对考生进行能力估计。但是，综合分析文献后发现，为数不多的未采用 Rasch 模型的计算机自适应语言测试研究在选择模型进行参数估计时，似乎也只是基于理论考虑或方便原则随机选取模型，并未考虑根据模型