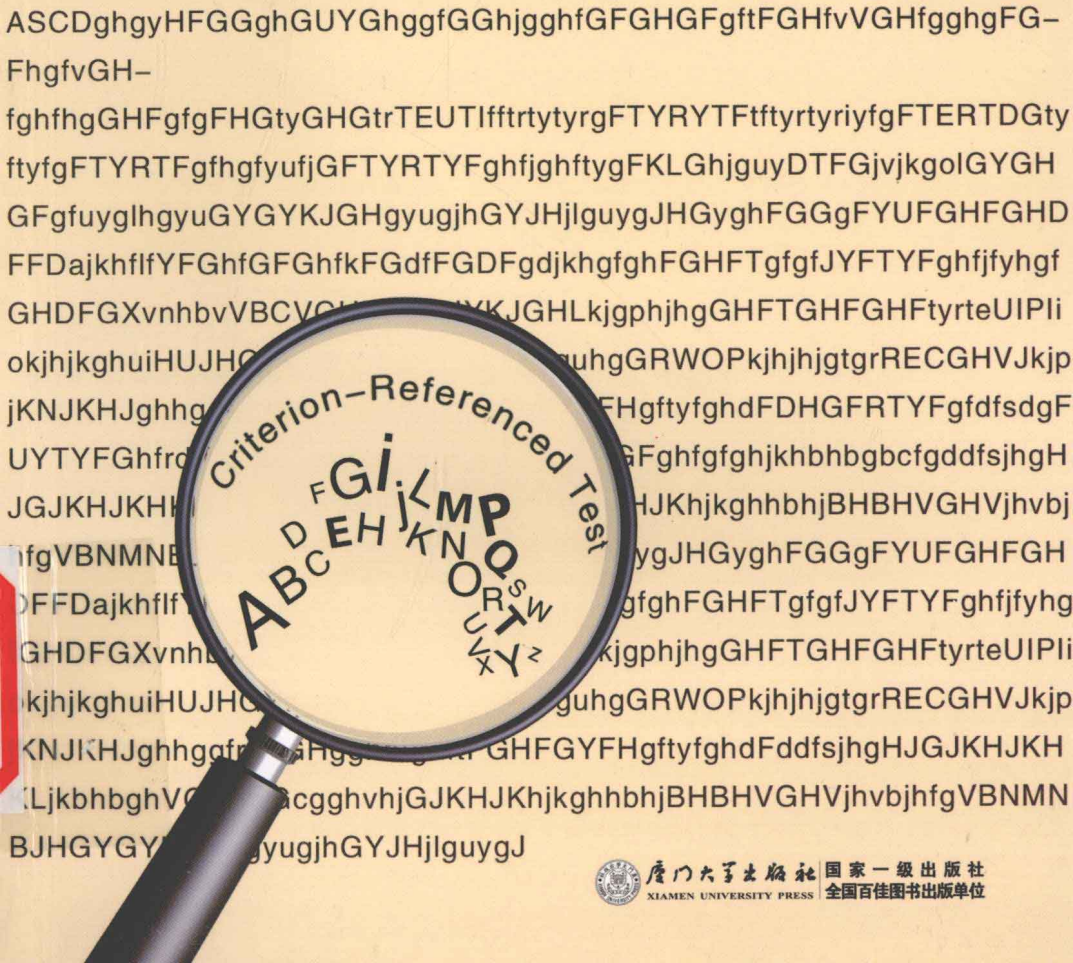
标准参照语言测试研究

Criterion-Referenced Test

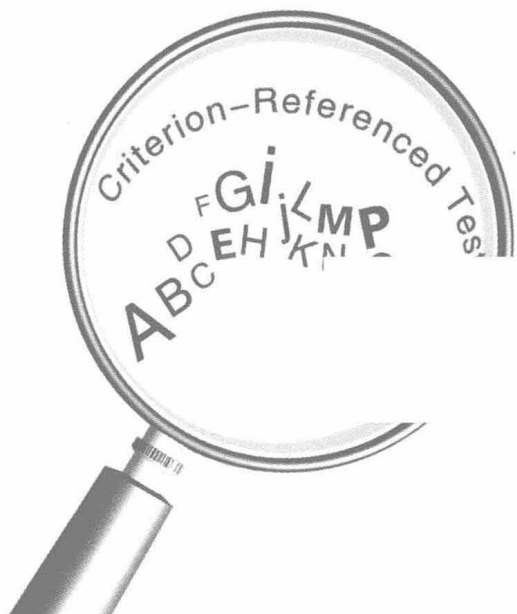
黄锐 著



厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

标准参照语言测试研究

黄锐 著



厦门大学出版社 XIAMEN UNIVERSITY PRESS
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

标准参照语言测试研究/黄锐著. —厦门:厦门大学出版社,2012.10
ISBN 978-7-5615-4456-3

Ⅰ.①标… Ⅱ.①黄… Ⅲ.①语言-测试-研究 Ⅳ.①H09

中国版本图书馆 CIP 数据核字(2012)第 248919 号

厦门大学出版社出版发行
(地址:厦门市软件园二期望海路 39 号 邮编:361008)
http://www.xmupress.com
xmup @ xmupress.com
厦门集大印刷厂印刷
2012 年 10 月第 1 版 2012 年 10 月第 1 次印刷
开本:889×1194 1/32 印张:9.75
字数:259 千字 印数:1～2 000 册
定价:36.00 元

本书获“集美大学优秀学术
著作出版基金”资助

序

黄锐的专著《标准参照语言测试研究》是在她的硕士学位论文《现代教育测量理论在标准参照语言测试中的应用与案例研究》的基础上，经过她多年的努力和孜孜不倦的耕耘、积累、拓展和深入研究，最终撰写而成的。该书现正式出版，作为她的老师，我感到颇为欣慰。

自从上个世纪60年代Glaser提出"标准参照测量"概念后，在教育测量界有关它的讨论就热门起来，国内对此的研究也在90年代迅速开展起来，到本世纪初达到高潮。中国汉语水平考试自1984年研制实施以来，语言测试界对标准参照测试的研究也随之发展，但就英语作为外语的语言测试界来说，对标准参照测试的研究却相对落后，国内目前尚无一本有关此内容的专著，研究该方向的文章也不多见。本书作者能在教育测量界和汉语测试界在标准参照测试研究的基础上，结合外语测试的特点撰写此书是一种大胆尝试。

本书从语言测试理论出发，通过与常模参照测试的理论和发展建构过程的比较，探讨了标准参照语言测试的理论和应用。信度和效度研究是语言测试中的难点，而标准参照测试的信度和效度更难把握。作者通过反复研读文献，以及几年的思考和写作，系统地梳理了标准参照语言测试的信度估计和效度研究的理论和方法，为同行们提供了一个比较全面了解标准参照语言测试全貌的专著。此外，作者还结合自身参与普通高等学校招生命题的工作经验，在书中就外语教师如何更好地参与命题、改卷、试题分析等方面的工作，提出了一些建设性意见。

该书对语言测试研究者、外语教师、高校师范生和语言测试的硕士生都有较高的参考和应用价值，是一本很值得阅读的专著。

是为序。

史秋衡

2012 年 9 月 21 日

于厦门大学颂恩楼

前 言

本书以语言测试的基本理论为基础，通过比较常模参照测试的理论和发展建构过程，探讨了标准参照语言测试的理论和应用，重点说明了标准参照语言测试的编制、项目分析以及标准参照语言测试的信度、效度和分数的解释。对标准参照语言测试的难点，即信度和效度进行了比较全面的探讨和研究，并通过理论联系实际的案例，采用了 NEATs 和 TEM 等国内大型的标准参照语言测试的统计数据，介绍了经典测量理论、概化理论和项目反应理论在标准参照语言测试的信度估计和效度验证中的应用，为同行学者提供了较多可参考和借鉴的依据。本书的主要读者对象为语言测试研究人员和外国语言学及应用语言学专业的硕士生、英语教师、英语专业的师范生以及对语言测试感兴趣的人士。

本书得以完稿和出版，首先要感谢我的授业恩师，教育部评估专家、国家“985 工程”创新基地副主任、厦门大学高等教育质量与评估研究所所长、教育研究院副院长、博士生导师史秋衡教授，尽管毕业多年，但他始终鼓励我在自己的研究领域里进行不断的探索和研究，并为本书作了序言；其次要感谢集美大学资助本书的出版并为我提供到广东外语外贸大学访学一年的机会，让我不但能全身心地完成书稿，还能直接地向语言测试界的专家们学习；感谢我的访学指导老师，广东外语外贸大学副校长、博士生导师刘建达教授和广东外语艺

术学院院长、博士生导师曾用强教授，是他们的教导和鼓励，让我能顺利地完成书稿；刘建达教授在百忙中通读了本书，并对书中的不足提出了宝贵的修改意见；广东外语外贸大学的博士生导师亓鲁霞教授对本书的目录和书名也提出过修改意见，在此一并表示感谢。

由于本人才疏学浅，书中难免有错误和不足，望读者批评指正。

黄　锐

2012 年 7 月于广东外语外贸大学

缩略语词汇

AERA	American Educational Research Association
APA	American Psychological Association
CI	Confidence Interval
CET 4 & 6	College English Test, Band 4&6
CRT	Criterion-referenced Test
CRLT	Criterion-referenced Language Test
CTT	Classical Testing Theory
DI	Difference Index
ETS	Educational Testing Service
ICC	Item Characteristic Curve
ID	Item Discrimination
IF	Item Facility
IELTS	The International English Language Testing System
IRT	Item Response Theory
IIFs	Item Information Functions
LAD	Language-acquisition Device
LID	Local Item Dependence
MLAT	Modern Language Aptitude Test
NCME	National Council on Measurement in Education
NEAT	National English Achievement Tests
NRT	Norm-referenced test
PETS	Public English Test System
PLAB	Pimsleur Language Aptitude Battery
SEM	Standard Error of the Measurement
TCC	Test Characteristic Curve
TEM 4 & 8	Test English for Major, Grade 4&8
TOEFL	The Test of English as a Foreign Language

目 录

第一章 绪 论

1.1 研究的目的和意义

中国教育心理界的权威人士张厚粲教授(1992)在她的《考试改革与标准参照测试》的开篇这样说道:"任何教育和心理测试的所得分数都必须有所参照才有实际意义,即测试分数只有与测试以外的某种指标进行比较,才能对应试者的分数给予适当的解释。"可见,分数的解释必须有个特定参照,才能全面地进行解释。在教育测量文献中,自 Glaser 在 20 世纪 60 年代提出"标准参照测量(Criterion-referenced measurement)"至今大概半个世纪的时间,标准参照测试对语言测试界来说仍然是新领域,因此,许多教师和考试管理者对其理论了解并不多,也不够深入。在国内的 CNKI 网站上,以"标准参照"和"语言测试"为关键词的文章,不足 5 篇,检索"标准参照"主题词,则多为教育或心理测量及医学类文献。故而,有必要从语言测试角度来探讨标准参照测试。尽管有学者认为要"告别标准参照测验和常模参照测验二元划分"(罗莲 2007),因为从同一测试得到的分数可作出常模参照和标准参照两种解释,两者是从分数解释的意义上划分的,并非两种不同的考试,但是,经过多年的实践和探讨,测量学界的研究者们基本上持两种不同的观点,绝大多数人认为,虽然标准参照和常模参照测试有一定区别,但两者是从不同角度对分数进行解释的。另外一个观点则认为,需要明确区分常模参照和标准参照测试,将标准参照测试的内容范围进行精确的定义后,其产生的分数可以进行常模参照的解释,但反过来是不行的(张凯 2002)。不管持哪种观点,我们知道,标准参照测试是在常模参照测试(Norm-

referenced Tests，NRT）的理论基础上发展起来的，可以认为是两种不同性质的考试，但又有许多相似之处。在应用标准参照测试的理论处理大规模考试时，又必须引入一些常模参照测试的办法，同时，在常模参照测试中，也借用标准参照测试中的一些思路来深入细致地分析题项和数据。两者互相借鉴和补充，从而丰富了测试理论。

上世纪80年代初，在我国影响和规模最大的普通高等学校招生全国统一考试中，引进了国外现代教育测量的理论和技术，对传统考试的办法进行了改革，形成了考试科学化、现代化的热潮。高考作为我国高等学校选拔新生的考试，最早它取用的是一种常模参照测试（张厚粲 1992），即学生从试卷上得到的成绩要和所有参加考试的考生成绩相比较，从而反映他在考生总体（常模总体）中的相对位置；随着教育改革的不断深化，考试改革已不仅局限在考试形式的改革，而是在考试内容、考试制度方面改革的深入发展，人们进一步要求完善考试的功能。随着高中新课改的不断深化，任何一名普通高中学生都需面对高中毕业会考，而这种旨在客观地评价中学生的学业情况，促进学生德、智、体、美、劳的全面发展，并用已定的标准为对照检验学生是否达到培养目标的考试，不再只是“常模参照测试”，而是标准参照测试（Criterion-referenced Tests，CRT）。如果参加会考的师生们还用常模参照测试中的理论和技术选定题项，他们就会发现很好的试题会被弃置不用。因此，靠单一的常模参照测试理论来选定题项、解释分数已远远不能满足当今日益完善的考试制度。

从国外教育测量史来看，20世纪初，在实验和教育统计的基础上形成了教育测量的理论，在30年代，教育测量的理论应用于选拔和人员安置的考试之中，到了四五十年代，这种着重于研究人员之间差异的常模参照性的测试运动达到了高潮。在这之后，随着教育、教学和教育测量研究的不断深入，人们已不满足于只用来反映考生做得好不好的测试结果，而是希望通过测试进一步了解考生在什么方面好，什么方面差以及优劣的程度，这样美国心理学家 R. Glaser 和 D. J. Klaus 首先于1962年提出了“标准参照测验”这一概念，其目的