

张艳莉 彭康洲 著

现代信息技术和语言测试研究： 方法与应用

Modern technology and language testing research:

Method and Application



北京师范大学出版集团

BEIJING NORMAL UNIVERSITY PUBLISHING GROUP

安徽大学出版社

现代信息技术和语言 测试研究：方法与应用

张艳莉 彭康洲 著



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
安徽大学出版社

图书在版编目(CIP)数据

现代信息技术和语言测试研究:方法与应用 / 张艳莉, 彭康洲著. —合肥:安徽大学出版社, 2012. 12

ISBN 978-7-5664-0606-4

I. ①现… II. ①张… ②彭… III. ①语言—测试—研究 IV. ①H09

中国版本图书馆 CIP 数据核字(2012)第 267326 号

上海外国语大学英语学科“211 工程”三期国家重点学科(英语)建设项目
(项目编号:211YYZYL01)

现代信息技术和语言测试研究:方法与应用 张艳莉 彭康洲 著

XIANDAIXINXIJI SHUHEYUYAN CESHIYANJIU; FANGFAYUYINGYONG

出版发行:北京师范大学出版集团

安徽大学出版社

(安徽省合肥市肥西路3号 邮编 230039)

www.bnupg.com.cn

www.ahupress.com.cn

印刷:合肥远东印务有限责任公司

经销:全国新华书店

开本:170mm×230mm

印张:10.75

字数:223千字

版次:2012年12月第1版

印次:2012年12月第1次印刷

定价:23.00元

ISBN 978-7-5664-0606-4

策划编辑:马晓波

责任编辑:朱楠 马晓波

责任校对:程中业

装帧设计:戴丽

美术编辑:戴丽

责任印制:陈如

版权所有 侵权必究

反盗版、侵权举报电话:0551—65106311

外埠邮购电话:0551—65107716

本书如有印装质量问题,请与印制管理部联系调换。

印制管理部电话:0551—65106311

前 言

自 20 世纪六七十年代语言测试成为独立的学科以来,经过多年的发展,在测试的研究方法、研究内容、测试因素及测试专业化研究方面有了长足进步。特别是近十几年来,基于各种先进的测量和测试理论的大量先进的统计软件被用于测试研究的方方面面,语言测试研究朝着越来越科学化、专业化的方向发展,人们对于了解测试的科学属性越来越有兴趣,同时对于正确判断考试质量、客观运用考试结果有了更多人文的思考。然而面对着专业化的术语和繁杂的软件应用,文科出身的研究者常常望而却步、踟蹰而行。鉴于此,我们在本书中介绍几种对语言测试研究非常重要的统计软件,分别是 SPSS、EQS、Bond&FoxSteps、Bond&FoxFacets 及 BILOG。通过介绍这几种软件操作的步骤,循序渐进,逐步讲解,采用图文并茂的形式,借助于软件应用实例的操作,对输出结果进行解释和分析,希望为广大教师和测试研究者提供参考。

本书共有六章。第一章通过对比国内外专业期刊中有关语言测试研究中的软件使用情况和研究内容,分析了国内和国际测试界在语言测试研究中的异同,探讨了现代信息技术在语言测试研究中的应用及发展趋势,以及对我国外语测试研究的启示。

第二章介绍了 SPSS 在语言测试中的运用,包括数据的输入与保存,及其在语言测试研究中的运用实例。SPSS 的运用实例中,我们选取了英语教学和英语考试实践中最为常见的案例,如期末考试成绩的描述统计分析(包括试题的平均分、标准差、方差、峰度和偏度等信息)、英语试卷的项目分析(包括题目的难度、区分度、信度等信息)、作文评阅的相关性和一致性分析、高考英语成绩与作文分数的线性回归分析、基于标准参照语言测试的期末考试听力理解的项目分析、职前教师 and 在职教师作文评阅比较、听力理解测试的因子分析等。

第三章介绍了结构方程模型软件 EQS 在语言测试中的运用,包括数据的

基本准备工作和研究举例。EQS的运用实例中,我们选取了常用的统计方法如线性回归分析和用于测试构念效度分析的验证性因子分析;其他的描述统计方法和信度分析方法基本和SPSS一致,在此不再赘言。

第四章和第五章分别介绍 Bond Fox(2007)开发的项目反应理论的分析软件 Bond&FoxSteps 和 Bond&FoxFacets 在语言测试研究中的应用,包括数据的输入与保存,以及应用举例;前者用于客观题分析,后者用于主观题分析。在客观题的研究案例中,我们介绍课堂测试中常见的单元测试的项目分析,包括题目是否存在性别差异的分析;主观题的研究案例中,我们介绍了教学质量评价中常见的教师说课比赛的分析情况。

第六章介绍了两分记法(正确为0,错误为1)的项目反应理论分析软件 BILOG-MG 在语言测试研究中的应用。在这一章,我们详细地介绍了项目的难度参数、区分度参数和猜测参数的统计方法和结果分析。同时,介绍了依据这三种参数进行题目筛选的方法。研究案例中,我们介绍了项目功能差异的诊断方法,审视测试题目的公平性问题。

本书只介绍语言测试研究中最常用的统计方法,无法将每种软件的所有功能都一一罗列。但我们在每章都附有使用了该软件的文章和专著,并作了简要介绍,对属于同一类型的文章或专著进行了归类,可供读者查阅,以弥补这一不足;同时提供这些软件的免费版本的下载地址和相关资源,希望对读者有所帮助。

由于著者能力有限,时间仓促,书中肯定有值得商榷的地方,欢迎同行批评指正。

著者

2012年9月

目 录

前言	1
第一章 信息技术在语言测试研究中的应用	1
1.1 前言	1
1.2 《语言测试》和国内重要期刊的英语测试研究对比	2
1.2.1 软件使用频率统计对比	2
1.2.2 研究内容对比	3
1.2.3 21 世纪国际和国内语言测试研究动向对比	7
1.2.4 国际和国内语言测试研究新进展	9
1.3 国际语言测试研究动向对我国外语测试研究的启示	11
1.4 结语	12
第二章 SPSS 在语言测试研究中的应用	13
2.1 数据的输入与保存	14
2.1.1 SPSS 界面	14
2.1.2 定义变量	15
2.1.3 输入数据	16
2.1.4 保存数据	17
2.2 SPSS 在语言测试研究中的应用举例	18
2.2.1 英语期末考试成绩的描述统计分析	18
2.2.2 英语试卷的项目分析	25
2.2.3 作文评阅的相关性和一致性分析	28

2.2.4	高考成绩与高考英语作文分数的线性回归分析	35
2.2.5	期末听力理解考试的项目分析	41
2.2.6	职前教师和在任教师作文评阅比较	44
2.2.7	听力理解测验的因子分析	48
2.3	运用 SPSS 软件分析的研究类文章和介绍 SPSS 的专著	56
第三章 EQS 在语言测试研究中的应用		59
3.1	数据的准备	61
3.1.1	EQS for Windows 界面	61
3.1.2	准备数据	62
3.2	EQS 在语言测试研究中的应用举例	63
3.2.1	英语学习风格与 TEM-4、TEM-8 成绩的线性回归分析	63
3.2.2	2007 年英语专业八级考试人文知识项目的验证性因子分析	68
3.3	运用 EQS 软件分析的研究类文章和介绍 EQS 的专著	75
第四章 Bond&FoxSteps 在语言测试研究中的应用		77
4.1	数据的输入与保存	78
4.1.1	Bond&FoxSteps 控制文件设置界面	78
4.1.2	控制文件设置	79
4.1.3	分析数据	81
4.1.4	查看输出结果	82
4.2	Bond&FoxSteps 在语言测试研究中的应用举例	89
	大学英语单元测验的项目功能偏差分析	89
第五章 Bond&FoxFacets 在语言测试研究中的应用		98
5.1	数据的输入与保存	98
5.1.1	Bond&FoxFacets 界面	98
5.1.2	Specification File 设置	99
5.1.3	分析数据	102
5.1.4	查看输出结果	102
5.2	Bond&FoxFacets 在语言测试研究中的应用举例	112
	英语专业教师说课比赛评分的公平性分析	112
5.3	运用 FACETS 软件分析的研究类文章和介绍 FACETS 的专著	126

第六章 BILOG—MG 在语言测试研究中的应用	129
6.1 BLM 控制句法文件的建立	130
6.1.1 BILOG—MG 界面	130
6.1.2 BLM 句法文件的建立	130
6.1.3 数据运行	135
6.1.4 查看输出结果	135
6.2 BILOG—MG 在语言测试研究中的应用举例	141
6.2.1 文理科学学生在测验中的项目功能差异研究	141
6.2.2 项目功能差异句法文件的建立	142
6.2.3 项目功能差异的输出结果	145
6.3 运用 BILOG 软件分析的研究类文章和关于项目反应理论的书籍	149
参考文献	150
附录	159
附录 1:Critical Values for Pearson Product Moment Correlation Coefficient	159
附录 2:Critical Values of the Spearman Rank Order Correlation Coefficients; The rs Tables	161
附录 3:Perceptual Learning—Style Preference Questionnaire	163

第一章 信息技术在语言测试研究中的应用

1.1 前言

20 世纪 60 年代,美国语言学家 Lado 的经典测试专著《语言测试》(Lado, 1961)成为语言测试学科的开山之作。Lado 以结构主义语言学为基础,认为语言能力可分,因此既可以分别测试听、说、读、写四种语言技能,也可以测试语音、句法、词汇、语法等语言知识。Lado 及其他语言测试专家将测量学和统计学的信度概念引入语言测试,标志着语言测试成为一门独立的学科。

创刊于 1984 年的专业学术期刊《语言测试》(Language Testing),大大推动了国际语言测试学科的发展。20 世纪 90 年代以来,语言测试在研究方法、测试内容、测试因素、真实性语言测试和测试的道德及专业化研究方面取得了长足进展。就定量方法而言,标准参照测量、概化理论、项目反应理论和结构方程建模在语言测试研究的各个层面得以广泛运用(Bachman, 2000)。因为,先进的研究手段不仅可以让我们关注考试分数,而且能更多地关注分数的意义和使用(Bachman, 2000: 25)。

外语教育工作者和研究者既有必要也有义务通过各种研究和统计手段了解学生的学习和考试情况,因为准确把握考试测量的正确性和有效性、甄别各种考试的质量有助于缓解考试给教师、学生和家长造成的压力。让我们了解并分析代表国外语言测试研究最新动态的《语言测试》期刊和国内重要期刊上刊登的有关语言测试论文中在过去的近 30 年中主要软件的使用情况,或许可以发现该领域最关心的研究话题和最常使用的研究手段和工具,必将会对我们的教学和测试研究工作有所启发。国内重要期刊以《中国学术期刊网络出版总

库》为准,选取了《外语界》、《中国应用语言学》、《外语教学与研究》、《外语电化教学》、《外语与外语教学》、《现代外语》、《外语教学理论与实践》、《外语教学》、《解放军外国语学院学报》、《现代教育技术》、《心理科学》、《外语测试与教学》等 12 种期刊上关于语言测试的实证研究类文章。

1.2 《语言测试》和国内重要期刊的英语测试研究对比

通过对《语言测试》期刊文章的检索,我们发现这些文章大致采用了 7 种统计软件:SPSS、EQS、LISREL、FACETS、GENOVA、WINSTEPS 和 BILOG,并以此为参照,统计国内语言测试论文使用以上统计软件的情况。就各软件的功能而言,SPSS 在社会科学领域使用最广泛,它包括了基本的描述统计、推断统计等等。EQS 和 LISREL 除拥有 SPSS 的基本功能以外,还可进行结构方程建模,处理多个原因和多个结果的关系,或者了解观察变量和潜在变量之间的关系。GENOVA 的基础是概化理论,用于研究多种误差方差的来源。FACETS 和 WINSTEPS 都是基于项目反应理论的单参数(难度系数)统计软件,分析考试成绩或者问卷调查的数学模型,前者主要应用于主观题,后者主要应用于客观题。BILOG 和前两种软件一样,也是建立在项目反应理论基础上的试题质量分析软件,用于两分记法(正确为 0,错误为 1)的试题分析,可以分析项目的难度、区分度和猜测度。

1.2.1 软件使用频率统计对比

自《语言测试》创刊以来,各统计软件使用的频率呈递增趋势(图 1.1)。SPSS 的使用频率最高,依次为 FACETS、EQS、GENOVA、BILOG 和 WINSTEPS。21 世纪前十年,FACETS 的使用频率增长最快,其后依次是 EQS、GENOVA、BILOG、SPSS,而 LISREL 的使用呈下降趋势。

从国内测试论文中使用的统计软件(图 1.2)来看,各软件的使用趋势并没有呈现与《语言测试》期刊中完全相同的规律。首先,SPSS 同样是国内研究者使用最多的统计工具,之后依次是 FACETS、LISREL、EQS、GENOVA 和 BILOG,这一规律和《语言测试》期刊基本吻合;但是,SPSS 在 20 世纪八九十年代的研究中使用相当少,而其他软件基本上还没有使用实例。进入 21 世纪以来,虽然 FACETS 和 LISREL 等软件在测试研究中有了一些使用案例,但 SPSS 仍然是国内研究者统计工具的首选,远远超过其他软件的总和。这与

SPSS 在国内的普及率不无关系:仅以当当网为例(截至 2012 年 9 月 17 日),有关 SPSS 的专著就多达 142 本,其中教材类 52 本;就外语教学研究而言,分别有王晓军(2006)《西部英语测试实证研究 & SPSS 操作》、秦晓晴(2004)《外语教学研究中的定量数据分析》和杨端和(2004)《语言研究应用 SPSS 软件实例大全》等多本专著。此外,有关结构方程模型的专著有 19 本,概化理论的专著有 13 本,项目反应理论的专著有 8 本,虽然这些统计软件涉及比较先进的测量理论,但在国内的语言测试研究中还没有得到广泛的运用。

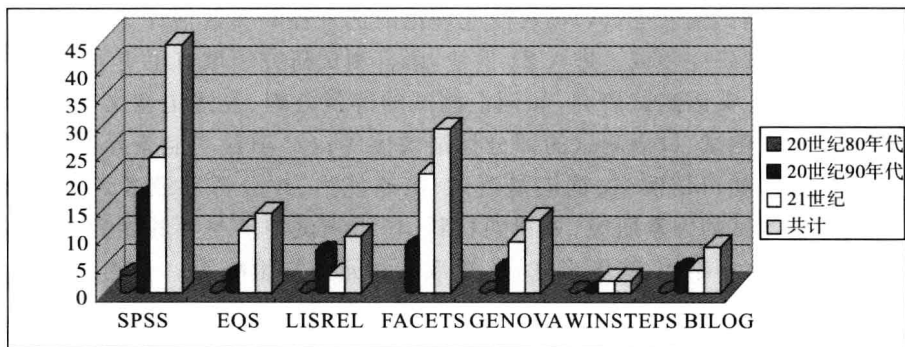


图 1.1 1984 年以来《语言测试》期刊文章中各软件使用情况统计

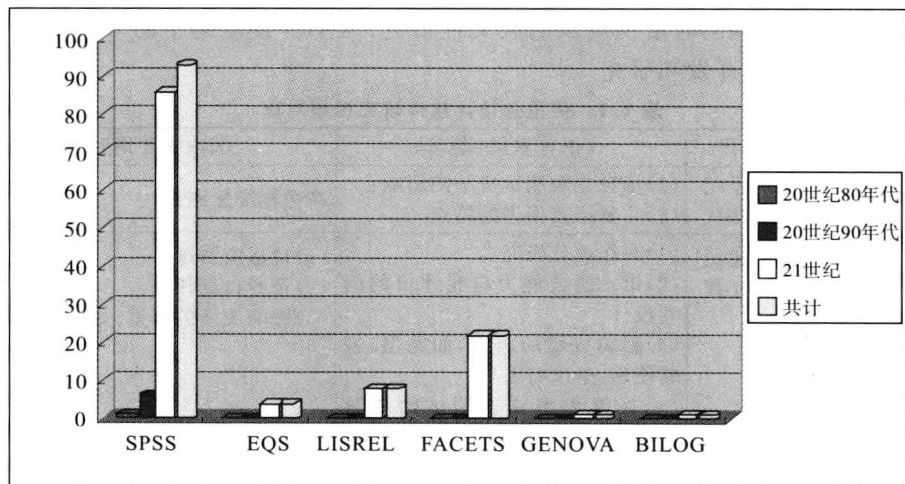


图 1.2 1984 年以来国内重要期刊中语言测试文章中各软件使用情况统计

1.2.2 研究内容对比

根据《语言测试》期刊中文章研究内容(表 1.1)和软件的使用情况,我们大

致可以了解国际语言测试研究的发展趋势。20 世纪 80 年代是《语言测试》期刊的起步阶段,语言测试的研究领域有限,使用软件作为定量研究工具的情况也比较少,仅有 SPSS 和 LISREL 两种,研究的范围大致涉及影响测试成绩的因素和构念效度,如完形填空、评分量表和语言能力自我评价研究等。20 世纪 90 年代是语言测试研究迅速发展的十年,主要涉及以下几个方面:(1)影响语言测试的各种因素,如题型、策略使用、学生特征;(2)信度研究,如标准参照语言测试、评分量表、评分员等因素;(3)效度研究,如缺词填空、不同测试群体的构念稳定性、内容效度、测试偏差等;(4)项目反应理论的适宜性和参数估计等,并首次进入少儿英语测试研究领域。进入 21 世纪,语言测试研究领域有了更大拓展,主要包括:(1)少儿英语测试研究,如词汇测试的等值处理、复述记叙文对叙事结构的把握、写作测试、口语测试的评分标准效度等;(2)项目功能差异研究;(3)测试效度研究,如口语测试、词汇量测试、策略能力、评分标准或者评分量表;(4)影响语言测试的因素研究,如视频材料、口语测试中参与者的能力、成员性格、文化图式、文本结构、答题形式、认知策略和元认知策略等;(5)测试偏差研究;(6)新题型的开发研究,如语法试题、读写综合等。从测试研究的整体趋势来看,随着测量理论和测试理论的发展,加上越加先进的统计软件的开发和应用,测试从重视客观性研究,过渡到重视主观性因素的研究;从只重视测试的科学性研究,过渡到同时重视测试的人文性研究;从单一技能题型的开发过渡到综合技能题型的开发和研究。

表 1.1 使用各统计软件研究话题对比

软件	年代	《语言测试》期刊	国内重要期刊
SPSS	20 世纪 80 年代	(1)场独立对测试成绩的影响 (2)儿童语言损害的特点	高考标准化测试
	20 世纪 90 年代	(1)写作评分研究 (2)第二语言能力自我评价的问卷设计 (3)影响成绩的因素,如题型、策略使用、学生特征 (4)标准参照语言测试的质量分析 (5)英汉翻译的质量研究 (6)基于测试任务特征的测试内容分析 (7)口语评分量表的制定 (8)缺词填空的效度	(1)自适应测试 (2)录音口试评分 (3)影响成绩的因素,如学习方法

续表

软件	年代	《语言测试》期刊	国内重要期刊
	21 世纪	(1) 配伍口语测试, 如成员性格、能力、语言背景 (2) 影响成绩的因素, 如视频材料、文化图式、文本结构、答题形式、认知策略和元认知策略 (3) 儿童英语测试 (4) 效度, 如答题过程、认知效度 (5) 以读促学, 读写综合 (6) 半直接口语测试的试卷形式和任务形式	(1) 反拨效应 (2) 影响成绩的因素, 如性格、写作环境、策略、焦虑、文章类型、题型、心理、性别、元认知策略、句法简化、工作记忆、模糊容忍度、体裁 (3) 计算机辅助考试和评分 (4) 效度, 如答题过程、能力构念、跨群体、理论效度 (5) 新型评价, 如自我评价、同伴互评、档案袋 (6) 考试方法, 如听写、朗读、成段改错、多媒体
EQS	20 世纪 90 年代	(1) 影响成绩的因素, 如策略使用和学生特征 (2) 效度, 如听力和阅读技能模型的跨组别稳定性	
	21 世纪	(1) 学生的语言能力和语言测试的构念 (2) 影响 SPEAK 图表描述任务的因素 (3) 测试任务特征 (4) 基于标准的课堂语言能力测试 (5) 阅读测试中的策略能力 (6) 口语测试评分量表 (7) 第二语言理解能力的可分性研究 (8) 儿童英语写作测试	(1) TEM-8 人文知识项目 (2) 职前中学英语教师口语考试 (3) 影响成绩的因素, 如考试方法效应、元认知策略 (4) TEM-4 阅读能力 (5) 自我效能感、学习焦虑和学习策略 (6) 元认知策略和认知策略 (7) 测试课程
LISREL	20 世纪 80 年代	(1) 第二语言能力的模型 (2) TOEFL 交际语言能力自评量表 (3) 完形填空	
	20 世纪 90 年代	(1) 专家评分员和新手评分员的作文评分信度研究 (2) 效度, 如少儿英语口语测试评分标准 (3) 分级测试	
	21 世纪	(1) 写作分数的概化系数 (2) 缺词填空的构念效度 (3) 托福网考的因子结构	(1) 学习动机和努力程度 (2) 写前计划变量 (3) 写作策略、写作认知过程

续表

软件	年代	《语言测试》期刊	国内重要期刊
FACETS	20 世纪 90 年代	(1)口语测试的偏差研究 (2)评分者训练效果研究 (3)语言能力量表 (4)口语评分的系统效应研究	
	21 世纪	(1)口语测试,如评分效应、性别偏差、小组讨论 (2)写作测试,如评分者偏差、评分员效应、自评、同伴互评、教师评价 (3)评分培训和评分反馈 (4)效度研究,如缺词填空、评分量表 (5)行为测试的分数线设置 (6)等值处理 (7)语法题的质量研究	(1)考试类型,如语用能力测试、话语填充测试、口语考试、写作、翻译测试、CET-6、TEM-8、非英语专业学位考试、读写综合 (2)评分员培训、评分员对比 (3)评分标准的制定
GENOVA	20 世纪 90 年代	(1)标准参照语言测试的信度 (2)口语测试的评分者信度 (3)内容效度专家判断法的一致性研究	
	21 世纪	(1)行为测试的分数线设置 (2)口语测试的评分者信度 (3)托福考试分数的可靠性指数 (4)写作测试的概化分数	口语考试
WINSTEPS	21 世纪	效度研究,如词汇量测试和口语测试	
BILOG	20 世纪 90 年代	(1)模型的适用性问题研究 (2)项目参数的估计	
	21 世纪	(1)口语测试的项目功能差异 (2)项目参数的估计 (3)儿童英语词汇测试的等值处理	听力项目的参数估计

从国内重要期刊上语言测试研究内容和使用的统计软件来看,可以大致得出国内语言测试研究的特点。20 世纪 80~90 年代是我国英语教学和测试的起步阶段,因此测试研究非常少见,仅涉及标准化考试、自适应测试试验,以及录音口试等;90 年代是国际语言测试界在研究方法、研究对象和统计手段等方面取得重大进展的阶段,可以说,这十年里国内语言测试研究与国际测试界有着明显的差距。进入 21 世纪,国内的语言测试研究取得了长足进展,出现了以 SPSS 运用为主的各项研究,具体体现在:(1)反拨效应;(2)影响测试的各种因

素,如学生性格、写作环境、策略、焦虑、文章类型、题型、心理、自我效能感、学习动机、努力程度等;(3)计算机辅助考试及评分;(4)新型评价方法,如自我评价、同伴互评、档案袋等;(5)考试方法,如听写、朗读、成段改错、多媒体、读写综合等;(6)各种考试效度和质量的研究,如 TEM-8 人文知识项目、TEM-4 阅读理解、职前中学英语教师口语考试等。另外,FACETS 等试题质量和效度的主观因素分析软件的应用,表明国内的测试研究者开始关注考试的公平性问题,如考试的主体尤其是学生和评卷人员对考试结果的影响。从整体趋势来看,国内的语言测试研究重点关注影响语言测试效度的各种因素,并逐步关心语言测试的公平性,这一研究趋势基本上和国际语言测试研究接轨。

从国际和国内语言测试研究采用的统计软件来看,我们大致可以得出以下结论。首先,SPSS 包含了统计分析中最基本的描述统计(平均分、标准差、方差、误差、正态分布)、相关分析、方差分析、回归分析、因子分析、信度等等,是语言测试研究中的必备工具。除 SPSS 以外,FACETS 是语言测试研究中使用最多的软件,在 21 世纪前十年被广泛运用;这体现了语言测试研究中对于主观评判方法的可靠性和有效性的重视,Bachman(2000:22)指出“我们不再争论项目反应理论模型的单维性和适用性问题,而是更多地关注其实用程度,尤其是借助它来研究涉及评卷人的做事测试(performance assessment,也叫真实性测试和评价)和计算机考试”。GENOVA 软件用来研究测试和评价中多种误差方差的来源,近几年的研究中,学者们常将它和 FACETS 结合起来分析数据(如 Bae & Bachman, 2010; Kozaki, 2004; Sawaki, 2007; Xi, 2010),这充分地印证了 Bachman(2000)的看法:21 世纪的很长时间里,我们将继续关注分数解释的有效性和考试使用带来的影响,研究方法不再是单一的定性研究或者定量研究,而是两种方法相互补充;研究工具不再是单一的工具,而是多种工具结合。EQS 用于分析多种因素间的关系,尤其是潜变量和观察变量的联系,在测试研究中的应用也较普遍;和它相类似的软件还有 AMOS、LISREL、Mplus 等,相比而言,EQS 易于操作。以 EQS 为主的结构方程模型分析软件将会在语言测试研究中继续得到应用,因为影响语言测试表现和成绩的因素还将是研究的重点之一,而这些都属于测试的效度研究范围。

1.2.3 21 世纪国际和国内语言测试研究动向对比

为了进一步了解进入 21 世纪以来国际语言测试界的研究动向,我们对比较分析了国际和国内 2005 年前后各软件的使用频率。

从国际语言测试研究所使用的统计软件(图 1.3)来看,2005 年前后各软件

的使用增长频率差异较大。从使用频率来讲,SPSS 还是使用最多的软件,但是其他软件的使用频率增长也很快,尤其是 FACETS、GENOVA、EQS 和 BILOG。语言测试的使用,正如 Bachman(2000)和 Bachman & Palmer(2010)所说是新世纪关注的重点,这一话题涉及考试分数所代表的具体含义,测试结果对考生语言能力的具体描述作用等等;考生和教师作为考试过程的重要利益群体,将会有更多的机会参与测试有效性的评判,这都需要研究评判的公正性,减少评判误差、了解误差来源等等,以上这些研究将会为以 FACETS 为首的项目反应理论软件(如 WINSTEPS、BILOG、Multi-LOG)和概化理论的软件 GENOVA 带来广泛的应用前景。

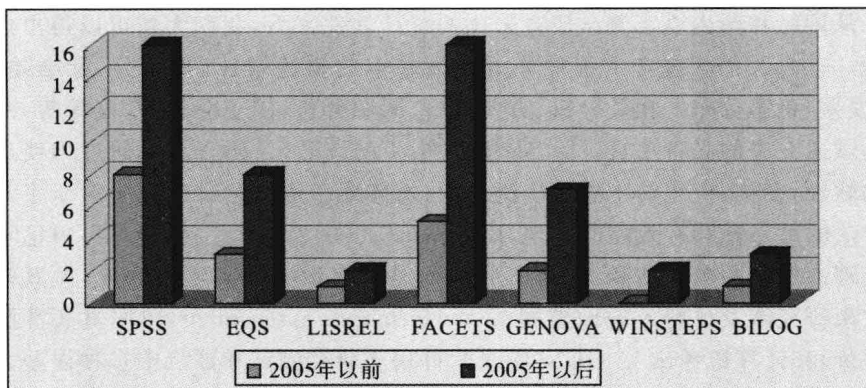


图 1.3 2000~2005 年和 2006~2011 年两个时间段《语言测试》文章中各软件的使用频率

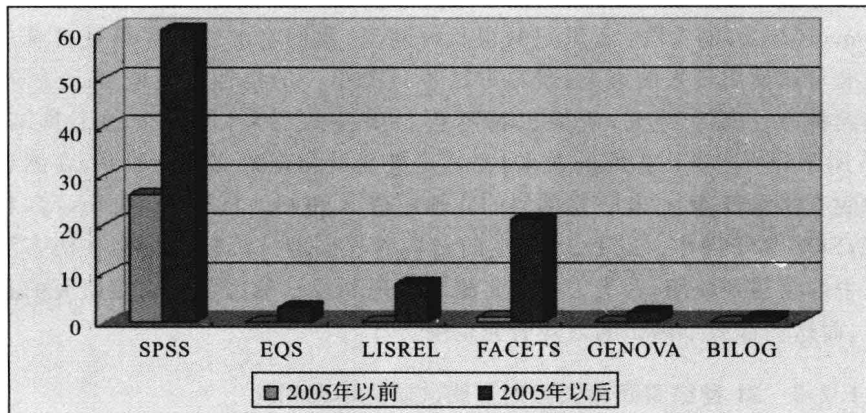


图 1.4 2000~2005 年和 2006~2011 年两个时间段国内语言测试文章中各软件的使用频率

从国内语言测试研究使用的统计软件(图 1.4)来看,SPSS 仍然占据主导地位,其他软件逐渐开始被语言测试研究者运用到测试及相关领域的研究中。

2005 年以前,统计软件完全以 SPSS 为主,2005 年以后的研究逐渐运用了项目反应理论、结构方程模型、概化理论等较为先进的测量理论。这说明,更多的国内的研究者在 2005 年以后开始关注教育测量理论中最先进的理论并将其运用到研究中。如前所述,国际国内的语言测试研究动向基本一致,虽然目前国内的研究整体上仍然数量有限。

1.2.4 国际和国内语言测试研究新进展

为深入了解国际语言测试研究的最新动向,我们将进一步对比《语言测试》期刊和国内重要期刊上最近三年(2009~2011)刊登的实证研究论文特别是涉及(口语测试和写作测试)研究话题,讨论其研究的深度。

从口语测试研究来看(表 1.2),《语言测试》期刊上的口语测试研究有 13 篇文章,大致包括六类:(1)影响口语测试成绩的因素,如线性图形描述任务、评分员对学生口音的熟悉度、考生的语言能力、本族语和非本族语的评分;(2)小组口语测试,如小组测试的考生人数和考生性格、有效互动的要素、互动的重要特征;(3)评分标准的制定方法;(4)口语能力的评价方法;(5)口语考试对课堂教学的反拨效应;(6)不同英语水平考生的口语特征。国内重要期刊上的口语测试研究有 7 篇文章,大致包括两类:(1)影响口语成绩的因素,如计算机辅助口语测试和面试,评卷人效应和偏差;(2)口试的效度和模型。整体来看,《语言测试》与国内重要期刊关注的重点不一样,国际语言测试界这几年比较注重新型口语测试研究,如小组测试,涉及影响因素和互动能力(interactional competence)的特征和构念。

表 1.2 近三年《语言测试》期刊和国内重要期刊上口语测试研究话题对比

《语言测试》	研究者和时间	国内重要期刊	研究者和时间
线性图形描述任务	Xi (2010)	机辅与面试	尹楠、郑玉荣、王丽丽、辛丹 (2010)
评分员对考生口音的熟悉度	Carey, Mannell & Dunn (2011)	计算机辅助口语测试	李玉平 (2009)
小组测试的考生人数和考生性格	Nakatsuhara (2011); Ockey (2009)	计算机考试评分者偏差分析	戴朝辉、尤其达 (2010)
小组测试的有效互动要素	Ducasse & Brown (2009)	口语考试的评卷人效应	刘建达 (2010)
小组测试互动的重要特征	May (2009)	口语考试模型	孙海洋 (2011)
考生的语言能力	Davies (2009)	口试的效度	文秋芳、王凌 (2009); 高怀勇、耿焕奇、刘锋、汪定明 (2011)