浙江理工大学学报，2020，44(3)：227-235
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673－3851(s).2020.03.001

国际语言测试研究的发展动态、热点及启示
——基于文献计量分析(2008—2019年)

金奚华
(浙江理工大学外国语学院，杭州 310018)

摘　要：为了把握国际语言测试研究的发展动态、热点和前沿趋势，借助文献计量软件CiteSpace，对Web of Science核心数据库收录的国际语言测试权威期刊*Language Assessment Quarterly*(《语言评测季刊》)和*Language Testing*(《语言测试》)2008—2019年发表的论文进行可视化知识图谱分析，从年度发文量、研究机构、高产作者、高影响力文献、高频关键词等方面进行了考察。通过文献关键词聚类分析发现了七大热点研究主题：学术写作(academic writing)、语言评价素养(language assessment literacy)、模型分析法(model)、专门用途医用英语测试(healthcare communication)、二语习得(second language acquisition)、评分员培训(rater training)和自动评分(automated scoring)，结合一些重要文献对热点研究主题分别进行了述评，进而提出该领域研究对本土语言测试研究的启示。

关键词：语言测试；CiteSpace；研究热点；综述；文献计量分析
中图分类号：H08　　**文献标志码**：A　　**文章编号**：1673-3851 (2020) 06-0227-09

Development trend, hot spots and enlightenment of international language testing research: Based on bibliometric analysis (2008—2019)

JIN Yanhua
(College of Foreign Languages, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In order to grasp the development trend, hotspots and cutting-edge trend of international language testing, this paper employs a bibliometric tool CiteSpace for visual knowledge mapping analysis of research articles published in authoritative international academic journals *Language Assessment Quarterly* and *Language Testing* from 2008 to 2019. The analysis is conducted from the following aspects: the number of articles published annually, research institution, high-yield author, literature with high influence, and high-frequency keyword. Seven hot research topics are found through keyword clustering analysis: academic writing, language assessment literacy, model, healthcare communication, second language acquisition, rater training, automated scoring, etc. The hot research topics are reviewed in combination of some important literatures, and then the enlightenment of the research in the field to native language testing research is proposed.

Key words: language testing; CiteSpace; hot research topics; review; bibliometric analysis

随着经济全球化的发展，外语教育作为一个国家的核心发展战略，需要不断健全教育体系和优化教育策略，以更好地服务于新时期国家的人才培养需求。2014年国家提出要加强外语测评体系的科学化、系统化建设，以促进外语教育教学的现代化发展，为学习者的语言能力发展提供连贯有序的阶

收稿日期：2020－02－12　　网络出版日期：2020－05－25
基金项目：浙江省哲学社会科学规划项目(20NDJC082YB)
作者简介：金奚华(1976－　)，女，浙江宁波人，讲师，硕士，主要从事语言测试学方面的研究。

梯[1]。了解国际语言测试领域的动态和热点研究，能更好借鉴国外成熟的理论和实证研究经验，促进本土语言测评研究。

自20世纪60年代以来，语言测试学界的理论和实证研究不断深化，学者们相继提出了“交际语言能力模型”[2]、“测试有用性框架”[3]，以及“基于证据的效验框架”[4]和“评测使用论证框架”[5]。近年来新的研究热点不断呈现，如认知诊断、评价素养等。学界对中国的语言测试研究进行了回顾，如蒋显菊[6]梳理了1996—2005年国内英语测试的研究，江进林[7]对2006—2017年外语测试实证研究进行了回顾。而对语言测试领域的国外研究综述主要集中在某一方面，如效度研究[8]、测试伦理[9]、作文自动评分[10]、语言测试中现代技术的应用等[11]以及质性研究语言测试热点[12]等。目前尚无学者对近十年来国际语言测试的发展动态、研究热点和前沿进行综述。在Web of Science(WoS)数据库中，*Language Assessment Quarterly*(《语言评测季刊》)和*Language Testing*(《语言测试》)为全球语言学SSCI收录的关于语言测试研究的国际权威期刊。周珊珊等[12]曾从研究方法、研究对象及研究问题三个方面对2011—2015年间这两本期刊的论文进行过质性研究。本文以这两本期刊的论文为样本来源，采用科学知识图谱分析方法，结合定性分析，系统梳理2008—2019年间的研究文献，并对其中的重要文献加以述评，通过分析国际语言测试研究的发展状况，希望为国内同行更深入地进行语言测试研究抛砖引玉。

一、数据来源与研究方法

(一)数据来源与数据收集

本文从WoS核心合集数据库中选择2008—2019年发表在*Language Testing*和*Language assessment quarterly*两本期刊的所有论文，剔除文献类型为Book Review、Correction、Editorial Material、Letter的文献，只保留Article类型文献，共得到476篇论文。建立原始数据文档，以纯文本格式保存所得论文的题录信息，包括题目、作者、作者单位、摘要、关键词、参考文献等。

(二)研究方法

知识图谱，又被称知识领域映射地图，能以可视化的方式显示知识发展进程与结构关系。将文献计量学的共被引分析、引文分析、词频分析等方法与信息可视化技术结合，通过数据挖掘、信息梳理和图形绘制，形象地展示学科的发展历史、核心结构、前沿领域及动态发展规律，从而为学科研究提供切实有价值的参考[13]2。本文采用文献计量分析工具CiteSpace软件对相关文献的共被引数据分析、关键词分析和聚类分析等，展示近十年来国际语言测试领域的知识基础、热点和研究趋势。

二、文献计量结果与分析

(一)发文量分析

国际语言测试研究领域的研究热度与该领域在权威期刊的文献发表数量、内容紧密相关。通过对*Language Testing*和*Language assessment quarterly*两本期刊2008—2019年的年度发文量变化，可以了解该领域的研究现状和发展情况，这两本期刊年度发文量具体如图1所示。

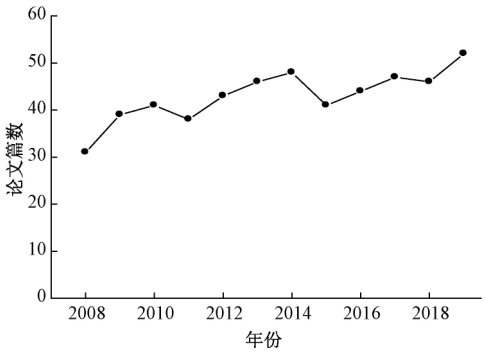

图1 2008—2019年期刊*Language Testing*和*Language assessment quarterly*年度发文量

通过两本期刊的发文量统计，发现2008—2019年该领域文献量总体稳中有升。在不同时段，文献发表增长的速度不同，分别于2010、2014、2017有阶段峰值，但达到高峰后逐渐回落，整体发文量呈现相对稳定的爬升状态。

(二)国家分布

本文用CiteSpace5.2软件，将节点类型选为“国家”，阈值调整为Top50 per slice即提取每个时间切片排名前50的数据来生成。2008—2019年两本期刊的语言测试领域论文作者所在的国家分布具体见图2，可见13个节点和20条链接线。其中年轮状的节点越大，国家字体越大，说明该国家作者发表论文的总体频次越高。连线表示国家间的作者有合作，连线越粗表示其共现频次越高。

从图2可以看出，不同国家之间普遍存在合作，排名前十的国家有美国、英国、澳大利亚、中国、加拿大、日本、荷兰、韩国、德国等。根据本研究统计，自

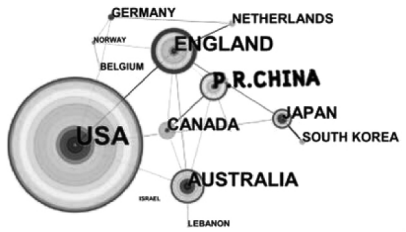

图 2　两本期刊发表语言测试领域论文的作者所在国家分布

2010 年来中国学者在 *Language Assessment Quarterly* 共发表 30 篇论文，在 *Language Testing* 发表了 10 篇论文，其中 2017 年为高峰期，在两个期刊共发表 8 篇论文。

(三)研究机构分布

本文将研究机构的图谱阈值设为每个时间切片前 50 项。发表语言测试领域论文的研究机构分布如图 3 所示，从图中可见 34 个节点和 12 条链接线。自 2008 年来，美国教育测验服务社(Educational Testing Service，ETS)是在这两本期刊发表论文最多的研究机构。ETS 举办的英语能力考试托福(TOEFL)，是海外多所院校外国学生申请入校学习的必要考试证明，因此其对托福考试相关的研究最多。其次是澳大利亚的墨尔本大学与美国密歇根州立大学。ETS、墨尔本大学和密歇根州立大学分别与多个研究机构合作密切。

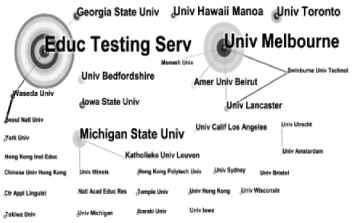

图 3　两本期刊发表语言测试领域论文的研究机构分布

(四)高影响力文献

文献的共被引分析(co-citation)指的是通过对一个文献空间数据集合进行文献共被引关系的挖掘过程。[13]139 共引文献数量越多，文献间的相关性越大。通过分析某篇文献同时被其他文献引用的频次高低，可以判断所在研究领域中具有重大影响的文献，即关键文献或核心文献。

本研究把节点类型设为“引用文献”，阈值设置为每个时间切片前 50 项、thresholding (c，cc，ccv)设置为(2，2，20)，图谱有 170 个节点和 304 条链接线。共被引文献字体大小与频次相关。年轮的颜色深浅与文献的时间和中心度相关；如浅灰色表示文献发表时间较早，灰色表示文献在较近的年份发表，黑色表示该文献有较大的中心度。连线表示共被引文献经常同时出现在一篇文献，连线粗细与共现频次相关。共被引高影响力文献分布如图 4 所示。

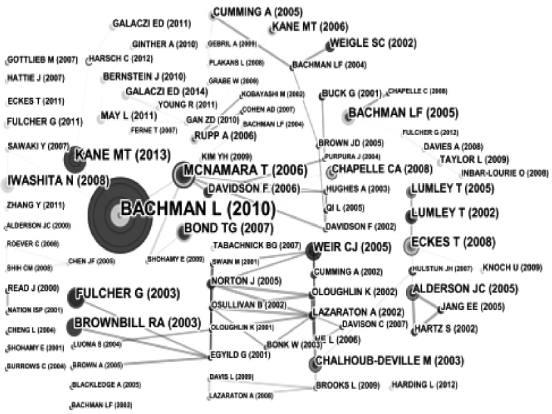

图 4　共被引高影响力文献知识图谱

选取其中共被引排名前十的文献，将其排名、被引频次、发表年份及作者制作成表，具体如表 1 所示。

表 1　高被引核心作者及其发表的文献

排名	被引频次	发表年份	第一作者	被引文献名
1	42	2010	Bachman L	*Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*
2	15	2013	Kane M T	*Validating the interpretations and uses of test scores*
3	14	2006	McNamara T	*Language Testing: The Social Dimension*
4	12	2003	Fulcher G	*Testing Second Language Speaking*
5	12	2007	Bond T G	*Applying the Rasch Model: Fundamental Measurement in the Human Sciences*
6	10	2008	Iwashita N	*Assessed levels of second language speaking proficiency: How distinct?*
7	10	2008	Eckes T	*Rater types in writing performance assessments: A classification approach to rater variability*
8	10	2005	Bachman L	*Building and supporting a case for test use*
9	9	2002	Lumley T	*Assessment criteria in a large-scale writing test: What do they really mean to the raters?*
10	9	2005	Weir C J	*Language Testing and Validation: An Evidence-Based Approach*

在上述核心文献中,Bachman的论著为语言测试的设计及有效论证奠定了重要的理论框架基础。在2005年之前,语言测试、教育和心理测量领域尚未制定出一套将考试成绩和基于分数的推论、考试使用以及考试使用后果联系起来的原则和程序。Bachman[14]提出如何构造测试使用,提供了从测试性能到解释以及从解释到使用的清晰链接。他提出了评测使用论证是将评测性能与使用(决策)联系起来的总体逻辑框架,其中评测使用论证将解释与决策联系起来,评测有效性论证将评估绩效与解释联系起来。在此基础上,Bachman等[5]发展了"评测使用论证"(Assessment use argument,AUA),该理论用来指导测试开发者如何科学地完成评测的设计并有效论证了其公平性的全过程。Bachman的著作[5]展现了语言评测领域在不同阶段的最新理念,并通过呈现完整的语言评测设计、开发和使用方法,对指导语言测试开发具有科学性和实用性的价值。

自1960年以来,效度研究一直是语言测试与评估的研究核心。效度验证理论以Kane的基于论证的效度验证模式和Weir的基于证据的效度研究最为突出。Kane[15]提出的基于论证的效度验证模式是一个完整的框架,其两个核心分别是解释论证和效度论证。解释论证指的是对分数和考试用途的解释,及根据受试考试中的表现推论出对分数的解释和基于考试所作出的决定这一过程的推论和假设;效度论证是在这一解释论证的框架下,收集证据并检验推论和假设是否合理的过程。该模式可以清楚地展现出证据搜集的先后顺序以及各种证据之间的内在联系,使其在效度验证时间中更具可操作性。Weir[4]认为效度研究需要不断地收集证据,从而论证测试结果的解释或使用是否具有合理性。效度证据来自于理论、环境、评分、校标关联、测试后果等5个方面。这两位作者的论著是语言测试领域关于效度理论与实践的重要文献。从社会维度研究语言测试是一个新兴的方向。

语言测试在社会中常用于身份鉴别、资格授予、控制移民等,如果其中存在不公平因素,会造成一定的社会负面影响。McNamara等[16]探讨了如何通过公平审查(fairness reviews)和道德规范(codes of ethics)来提高语言测试的公平性。由于测试内容和测试方式会受价值导向和政府政策的影响,语言测试在教育系统中的使用同样具有社会性。因此,语言测试在社会中的使用和影响与测试的信度、效度等技术指标同等重要,需要引起关注。该文献对语言测试社会维度进行了全面而深入的探讨,是学界对语言测试进行跨学科研究的新起点。

口语测试由于具有灵活多变、主观性强等特点,加之主考教师把握测试评价标准存在客观差异,对其信度、效度和公平度的研究存在相当的难度。Fulcher[17]梳理了口语测试历史、并从理论和实践研究的角度,为二语口试提供了更加完善的理论体系和切实可行的评价标准,成为近几年口语测试研究的重要的文献之一,其口语测试描述任务框架(a framework for describing tasks)包括口语测试设计过程、测试任务的种类及评价等。而Iwashita等[18]开展了针对英语作为第二语言的口语能力的性质调查,在开发新托福国际考试的评分量表的需求下,他们研究了考生口语的详细特征与评分者对这些成绩的整体评分之间的关系。该项研究通过对口语熟练程度的本质进行洞察,用语言等级表来衡量口语熟练程度,对二语习得的语言测试中使用恰当性的方法产生了重要影响。

有关评分员的评分过程方面的实证文献,对后续相关研究具有一定的借鉴价值。广受关注的Eckes[19]通过多面Rasch模型,分析了写作评分员的六种类型、存在的明显评分特征及评分模式差异。该文献对评估有评分员介入的大规模语言测试质量、评分员监测和评分员培训均具有重要的指导意义。而Lumley[20]的实证研究是基于澳大利亚政府用于协助移民决策的英语水平特殊测试(STEP)。他用评分员提供评分的有声思维法,对评分员在分析评级量表中的评分类别以及评级员在评级过程中面临的困难进行了解释。该文献为评分员如何对二语学习者在书面语言评分量表做出评分决定的过程进行了有效探索。

近十年,语言测试研究方法中引起关注的重要文献还有Bond等[21]的著作《应用rasch模型:人文学科的基本测量》,其中讨论了Rasch模型的基本原理,及其在人文学科各领域的应用。Rasch模型是一个著名的心理统计学模型,是丹麦数学家Georg Rasch在项目反应模型基础上发展的一个潜在特质模型。Rasch模型的主要特点是个体与题目共用标尺、线性数据、参数分离等,能够确保客观测量的实现,在国外学术界受到广泛关注和深入研究[22-23]。近年来,国内语言测试界应用项目反应理论和Rasch模型也进行了很多实证性研究,包括将其运用在测试等值[24]、测试信效度[25]、题库建设[26]等领域。

科学知识具有累积性、连续性和继承性的特征[27],科研文献也不是孤立的,而是相互联系、不断延伸的。核心文献在科学知识系统中起到了关键链

接和导引的作用，将导引学者对某个研究领域进行快速、准确定位。图 5 呈现了十年来语言测试研究中非常重要的理论基础，如测试开发、效度验证模式、基于社会维度的语言测试、口语测试框架等。热点文献还呈现了基于 Rasch 模型的研究方法在语言测试领域受到了广泛的引用，而且引领了评分员研究、口语测试等领域的实证类研究。

(五)研究热点可视化图谱分析

关键词代表了文章的核心内容，如果特定时期关键词在某一领域的文献中反复出现，就可以将其视为这个时期的研究热点[13]200。本研究在 CiteSpace5. 2 中设置 Top 50per slice，thresholding (c，cc，ccv)为(2，2，20)，最后得到 60 个节点，73 条链接线。语言测试文献关键词分布见图 5，其中展示了 2008—2019 年国际语言测试界的研究热点。本研究将这些关键词的主题和属性进行聚类分析，发现以下七个热点研究主题：学术写作(academic writing)、语言评价素养(language assessment literacy)、模型分析法(model)、专门用途医用英语测试(healthcare communication)、二语习得(second language acquisition)、评分员培训(rater training)和自动评分(automated scoring)。文献关键词聚类图如图 6 所示。

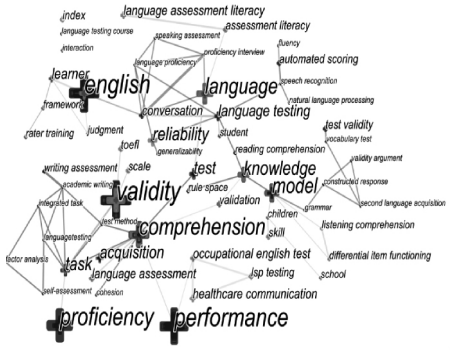

图 5　语言测试文献关键词图谱

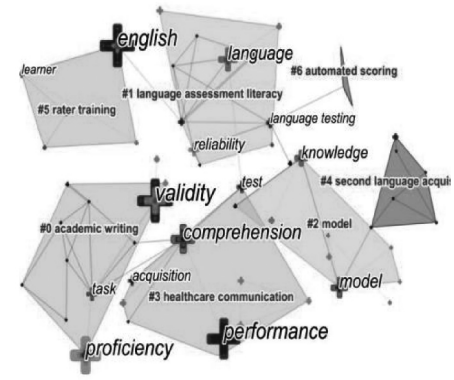

图 6　语言测试文献关键词聚类图谱

下面依次对热点主题及相关的文献研究展开述评。

a)学术写作(academic writing)：从 Barber 关于科技论文可测量语言特征的研究开始，国外学术写作研究至今已有 50 多年的历史[28]。由于母语环境和二语环境存在差异，而且二语作者类型特征不同，使二语学术写作研究复杂化，需要大量的实证研究。研究者从学术写作能力和二语水平之间的关系、影响二语学术写作水平变化的因素、二语学术写作过程以及文本特征等多方面探讨了如何发展学术写作能力。如 Appel 等[29]调查了加拿大学术英语评估(Canadian Academic English Language Assessment，CAEL)中不同水平的二语学习者在其学术论文中重复使用英语词汇组合的情况。研究发现，写作水平低的作者对测试材料的阅读文章依赖性较高，更多使用了原文的立场和话语组织表达语句。Llosa 等[30]以美国大学 103 名国际非英语母语本科生为研究对象，调查学生在完成托福写作任务和美国大学必修写作课程中的表现，发现托福任务成绩(独立、综合、写作总分)与教师对学生整体英语水平和写作水平的评价呈中度显著相关。Barkaoui[31]探讨了参加 PTE (Pearson test of English)学术考试的考生第二语言(L2)写作成绩变异的来源，研究发现，考生的初始英语语言能力对 PTE 学术写作成绩呈最强预测。

b)语言评价素养(language assessment literacy)：语言评价素养认为外语教师要培养对测试和评价原则和实践的理解[32]，包括设计、命制、修改和评价大规模标注化考试和课堂测验所需的知识、技能和能力，并需要熟悉测试过程[33]。近年来的研究包括语言教师对评价素养培训需求调查，培训对教师相关知识积累发挥作用。如，Vogt 等[34]使用问卷调查，并访谈了来自 7 个欧洲国家的 63 位教师，评估了目前外语教师的语言评价素养水平，并确定他们在这一领域的培训需求。Lam[35]通过对香港中小学教育机构的调查，发现香港的语言评估培训仍然不足，进而提出了如何提高语言评估素养的建议。Baker 等[36]探讨了海地教师与语言评估专家在培训及合作过程中的语言评估素养发展。

c)模型分析法(model)：在语言测试研究中，基于 Rasch 模型的实证研究已成为主流方法。单参数 Rasch 模型能同时评估项目难度和考生能力；多层面 Rasch 模型(Multi-faceted Rasch Model，MFRM)除了能估计项目难度和考生能力外，还能分析其它多个层面，并可以对各个层面的交互效应、数据对模型的拟合程度等进行计算。近年来，尤其

是MFRM被广泛应用到各类测试的影响因素分析,如听力测试视听通道的研究[37]、写作测试任务变量的研究[38]等。应用最多的是对评分员效应的研究[22,39-40]。如,以托福国际商务英语口语考试的成绩为样本,Davis[23]应用MFRM研究了评分员培训和经验对口语评分成绩的影响。Trace等[40]通过MFRM分析了评分员协商对第二语言写作评分过程中评分员的严厉度、自我一致性及偏见的影响。

d)专门用途医用英语测试(healthcare communication):由于国家和地区间人口流动和社会分工日益精细,用于专业目的的语言评测成为焦点[41]。专门用途英语测试的研究可以减低一些大规模测试的社会风险,因此需要提高专业化语言水平测试的准确性。近几年关于医疗交流的测试比较热门,相关研究主要有行业评分标准的建构、测试任务的真实性等。例如,O'hagan等[42]将医疗行业的评分标准纳入职业英语测试(Occupational English Test, OET)的口语测试结构,研究发现新标准的评分具有较高的信度,新的和现有的测试标准代表的构念也基本一致。结果表明,纳入行业评分标准扩展专门用途英语测试的结构具有可行性。Woodward-Kron等[43]从语篇分析角度探讨了OET口语部分的任务真实性。通过比较国际医学毕业生和准医生在任务中的语篇结构和语言表现,该研究提出,应提高职业英语测试的性能,以提高更多与职业相关的语言交流技能。

e)二语习得(second language acquisition):语言测试与二语习得研究的关系密切,通过语言测试可以检验二语习得的过程和方式、检验二语习得研究者所提出的假设。近年来在二语的词汇习得、听力学习、写作能力发展方式、阅读认知诊断等方面都有相关研究。例如:Wigglesworth等[44]在二语写作课堂上评估比较了成对的学习者产生的文本与个体学习者产生的文本,研究显示,合作写作的学习方式对二语学习者议论文写作准确性有正面影响,但对写作的流利性和复杂性没有影响。Wagner[45]通过设计二语听力测试的不同方式,考查其对考生成绩的相关性,并探讨了在二语听力学习过程中视频技术辅助对学习者的影响。Kim[46]探讨了如何在测试中对新生进行二语阅读能力的认知诊断评估。通过专家对试卷的测试题目进行编码,用来诊断学生的知识、技能、阅读策略等,评估的信息不仅给学生个体反馈了其二语阅读过程中的优缺点,而且有助于相关管理者和教师提升阅读课程教学及开发教学材料。

f)评分员培训(rater training):评分员在口语和写作评价中具有重要地位。评分员变量复杂,包括个人经验、背景及认知等,因此在评分过程中会引入与测试构念无关的因素,从而影响评分的效度[38,47-48]。近年来,涌现了很多对评分员的培训研究,如何通过有效的培训减少评分员个人因素带来的评分差异成为研究热点。例如,Lim[49]研究了新手和有经验评分员在12~21个月多个时间点上的纵向评分表现,结果发现,部分新手评分员在最初表现不同的情况下,相对较快地学会了适当的评级;评分员能够在研究时间段内保持评分质量;评阅量和评分质量相关。Kim[50]研究了培训对新手、成长型和老手评分员的评分变化,结果发现老手评分员保持评分稳定发挥,新手评分员虽然经历培训但仍存在问题,成长型评分员进步最大。

g)自动评分(automated scoring):随着社会科学技术的进步进而人工智能的发展,计算机自动评分逐渐成为研究热点。计算机自动评分具有可靠、客观、节省人力和物力等优势特征,而且其即时性和互动性优于人工评分[51]。目前,计算机自动作文评分(Automated Essay Scoring)不但采用统计、自然语言处理以及人工智能等方面最新成果对作文进行评分,而且也可以完成开放性问答题(如阅读理解中的简答题)的评分[52]。此外,随着语音识别技术的应用,计算机自动评分在口语考试领域也逐渐开展[44,53]。自动评分的信、效度一直是关注的焦点,尤其在主观题评阅上,计算机自动评分与人工评分的差异还需要大量的实证研究。Giang等[54]比较了人工评分员和自动论文评估工具(Automated essay evaluation,AEE),发现AEE与人类评分只有适度的相关性,局限在于AEE只能识别内容词而不是语篇层面上的组织方式,因此检测不到稍微偏离主题的文章和抄袭。Kang等[55]评估了一套计算机算法,以剑桥英语语言测试(Cambridge English Language Assessment,CELA)对二语学习者的英语独白进行的评估为研究对象,发现计算机与官方CELA计算结果显著相关。

三、总结与启示

本文基于WoS核心数据库,借助文献计量分析工具CiteSpace,对2008—2019年发表在国际语言测试研究的权威期刊*Language Assessment Quarterly*(《语言评测季刊》)和*Language Testing*(《语言测试》)的476篇文献进行研究,梳理了发文

量、研究机构、高影响力文献和研究热点等方面。研究发现：2008—2019 年国际语言测试研究文献发表量整体稳中有升，研究机构及不同国家之间的合作较为普遍。通过高影响力的关键文献图谱发现，语言测试指导理论包括“评测使用论证”、效度验证模式等，把握了口语测评、评分员研究、基于 Rasch 模型的语言测试研究等新兴方向。研究借助图谱对文献关键词的主题和属性进行聚类，得出七个热点主题：学术写作、语言评价素养、模型分析法、专门用途医用英语测试、二语习得、评分员培训和自动评分。与国际热点研究主题相比较，近十年国内语言测试的实证研究虽然涉及到写作测试、Rasch 模型的应用、评分员培训等方面[7,24-26]，但在某些测试研究新兴领域，本土研究尚处于起步阶段。因此从本研究可以得到如下启示：

其一，随着科技日新月异，人工智能应用迅速发展并影响着国际教育领域。在国际语言测试研究中，热点主题如写作自动评分、口语的人工智能评分、计算机自适应测试等为外语测试现代化提供了新方法和新思路。本土研究可以借鉴先进技术和研究经验，改变语言测试的形式和手段，提高语言测试的信效度，从而更精准反馈学生的学习情况，掌握学生的学习能力进展进程，使语言测试更好服务于教育教学的各种需求。

其二，语言测评对外语教学具有重要的反拨功能，教师的评价素养与教学质量紧密相关。在国际语言测试研究中评价素养是新兴热点主题。在国内，一线教师基本没有经过相关语言测试理论等培训，在评价素养研究方面接近空白。因此，可加强该研究主题的交流，关注教师评价素养的培训并进行实践研究。在教学中，教师可以更好地利用评价手段收集学生的学习信息，并正确利用所得的信息进行教学决策和改进，改变不当评价的误用和由此导致的偏见，更好地引导学生建立学习目标，关注学习进展并参与评价建设等，从而推动本土外语教学更好地发展。

其三，专门用途英语测试是针对不同职业能力与英语测试标准的对接。该领域国际语言测试的研究热点近几年以医疗行业比较突出。在全球化的发展背景下，国家的许多行业人才需要具备国际化的视野、熟练的外语水平和高素质的综合能力。而职业的差异对语言能力的要求不同，语言的评价方式需结合不同职业的特点。目前，本土外语测试主要研究对象以大学生为主，而如何使英语水平测试标准与不同职业能力相对接研究尚待开发。因此外语测试研究不仅直接作用于教育教学，而且还需要面向就业需求为外语类人才建设作出贡献。

其四，从发文作者所在国家的聚类共现图谱可知，语言测试研究的国际合作普遍存在。以《中国能力等级量表》为例，现教育部考试中心已与英国文化教育协会，及剑桥大学英语考评部研究雅思、普思考试与量表的对接，同时教育部与美国教育考试服务中心对托福与量表的对接也展开了研究，其研究成果将会对成绩使用机构、学校教师、学生，乃至国家外语教育改革等产生多方面的影响。语言测试研究必然走向国际化，研究者应具有国际视野，结合本国有特色的语言测试研究，融合创新，更新语言测试理念，改革考试内容，增加创新题型，不断促进外语教育教学改革。

参考文献：

[1] 中华人民共和国国务院.国务院关于深化考试招生制度改革的实施意见[A/OL].(2014-09-04)[2020-02-17].http://www.gov.cn/zhengce/content/2014-09/04/content_9065.htm.

[2] Bachman L F.Fundamental Considerations in Language Testing[M].Oxford：Oxford University Press，1990：111-159.

[3] Bachman L F，Palmer A S.Language Testing in Practice[M].Oxford：Oxford University Press，1996：85-231.

[4] Weir C J.Language Testing and Validation：An Evidence-Based Approach [M].New York：Palgrave Macmilla.2005：85-108.

[5] Bachman L F，Palmer A S.Language Assessment in Practice：Developing Language Assessments and Justifying Their Use in the Real World [M].Oxford：Oxford University Press，2010：75-130.

[6] 蒋显菊.国内英语测试研究：十年回顾与展望[J].外语界，2007(2)：89-96.

[7] 江进林.我国外语测试实证研究：回顾与展望：基于外语类主要期刊的统计分析(2006—2017)[J].外语界，2018，185(2)：40-48.

[8] 李清华.语言测试之效度理论发展五十年[J].现代外语，2006，29 (1)：87-95.

[9] 甘凌，夏纪梅.语言测试伦理问题研究：回顾与启示[J].广东外语外贸大学学报，2016，27 (2)：59-64.

[10] 梁茂成，文秋芳.国外作文自动评分系统评述及启示[J].外语电化教学，2007(5)：18-24.

[11] 刘建达.现代技术与语言测试：应用、影响及发展方向[J].外语电化教学，2013(3)：46-51.

[12] 周珊珊,罗少茜,赵海永.国外语言测试研究热点综述(2011—2015年)[J].外语测试与教学,2018(2):1-14.
[13] 李杰,陈超美. CiteSpace:科技文本挖掘及可视化[M].北京:首都经济贸易大学出版社,2017.
[14] Bachman L F. Building and supporting a case for test use[J]. Language Assessment Quarterly, 2005, 2 (1): 1-34.
[15] Kane M T. Validating the interpretations and uses of test scores[J]. Journal of Educational Measurement, 2013, 50 (1): 1-73.
[16] McNamara T, Roever C. Language Testing: The Social Dimension [M]. Oxford: Blackwell Publishing Ltd, 2006: 129-148.
[17] Fulcher G. Testing Second Language Speaking[M]. London: Pearson Education Ltd,2003:171-198.
[18] Iwashita N, Brown A, McNamara T, et al. Assessed Levels of second language speaking proficiency: How distinct?[J]. Applied Linguistics, 2008, 29 (1): 24-49.
[19] Eckes T. Rater types in writing performance assessments: A classification approach to rater variability[J]. Language Testing, 2008, 25 (2): 155-185.
[20] Lumley T. Assessment criteria in a large-scale writing test: What do they really mean to the raters?[J]. Language Testing, 2002, 19 (3): 246-276.
[21] Bond T G, Fox C M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences [M]. New York: Routledge,2007: 31-51.
[22] Schaefer E. Rater bias patterns in an EFL writing assessment[J]. Language Testing, 2008, 25 (4): 465-493.
[23] Davis L. The influence of training and experience on rater performance in scoring spoken language [J]. Language Testing, 2016, 33 (1): 117-135.
[24] 刘建达,吕剑涛. Rasch模型等值多套英语试卷的可行性研究[J].现代外语,2012(4):401-408.
[25] 江进林,文秋芳.基于Rasch模型的翻译测试效度研究[J].外语电化教学,2010(1):14-18.
[26] 杨志明.考试公平性之题目及试卷功能差异探析[J].教育测量与评价:理论版,2017(9):5-12.
[27] 尹春丽.科学学引文网络的结构研究[D].大连:大连理工大学,2006:25-50.
[28] 毕劲,秦晓晴,宋德伟.国外二语学术写作研究趋势及其启示[J].外语教学,2014,35(2):45-49.
[29] Appel R, Wood D. Recurrent word combinations in EAP test-taker writing: Differences between high- and low-proficiency levels [J]. Language Assessment Quarterly, 2016, 13 (1): 55-71.
[30] Llosa L, Malone M E. Comparability of students' writing performance on TOEFL iBT and in required university writing courses [J]. Language Testing, 2019, 36 (2): 235-263.
[31] Barkaoui K. Examining sources of variability in repeaters' L2 writing scores: The case of the PTE academic writing section[J]. Language Testing, 2019, 36 (1): 3-25.
[32] Boyles P. Assessment literacy[A]// Rosenbusch M (ed.). National Assessment Summit Papers[C]. Ames I A: Iowa State University, 2005:11-15.
[33] Fulcher G. Assessment literacy for the language classroom[J]. Language Assessment Quarterly, 2012, 9 (2): 113-132.
[34] Vogt K, Tsagari D. Assessment literacy of foreign language teachers: Findings of a European study[J]. Language Assessment Quarterly, 2014, 11(4): 374-402.
[35] Lam R. Language assessment training in Hong Kong: Implications for language assessment literacy [J]. Language Testing, 2015, 32 (2): 169-197.
[36] Baker B A, Riches C. The development of EFL examinations in Haiti: Collaboration and language assessment literacy development [J]. Language Testing, 2018, 35 (4): 557-581.
[37] Batty A O. A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning [J]. Language Testing, 2015, 32 (1): 3-20.
[38] Li J L. Establishing comparability across writing tasks with picture prompts of three alternate tests [J]. Language Assessment Quarterly, 2018, 15 (4): 368-386.
[39] Eckes T. Operational rater types in writing assessments: A classification approach to rater variability[J]. Language Assessment Quarterly, 2012, 9 (3): 270-292.
[40] Trace J, Janssen G, Meier V. Measuring the impact of rater negotiation in writing performance assessment [J]. Language Testing, 2017, 34 (1): 3-22.
[41] Shohamy E, Hornberger N. Language Testing and Assessment[M]. 2nd Eds,New York: Springer,2008: 397-416.
[42] O'hagan S, Pill J, Zhang Y. Extending the scope of speaking assessment criteria in a specific-purpose language test: Operationalizing a health professional perspective[J]. Language Testing, 2016, 33 (2): 195-

216.

[43] Woodward-kron R, Elder C. A comparative discourse study of simulated clinical roleplays in two assessment contexts: Validating a specific-purpose language test [J]. Language Testing, 2016, 33 (2): 251-270.

[44] Wigglesworth G, Storch N. Pair versus individual writing: Effects on fluency, complexity and accuracy [J]. Language Testing, 2009, 26 (3): 445-466.

[45] Wagner E. An investigation of how the channel of input and access to test questions affect L2 listening test performance [J]. Language Assessment Quarterly, 2013, 10 (2): 178-195.

[46] Kim Ah-Young. Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability [J]. Language Testing, 2015, 32 (2): 227-258.

[47] Wei J, Llosa L. Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks [J]. Language Assessment Quarterly, 2015, 12 (3): 283-304.

[48] Winke P, Gass S, Myford C. Rater's L2 background as a potential source of bias in rating oral performance [J]. Language Testing, 2013, 30 (2): 231-252.

[49] Lim G S. The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters [J]. Language Testing, 2011, 28 (4): 543-560.

[50] Kim H J. A qualitative analysis of rater behavior on an L2 speaking assessment [J]. Language Assessment Quarterly, 2015, 12 (3): 239-261.

[51] Landauer T K, Laham D, Foltz P W. Automated scoring and annotation of essays with the intelligent essay assessor [C]//Shermis M D, Burstein J. Automated Essay Scoring: A Cross-Disciplinary Perspective. Mahwah N J. Lawrence Erlbaum Associates, 2003: 87-112.

[52] Carr N T, Xi X. Automated scoring of short-answer reading items: Implications for constructs [J]. Language Assessment Quarterly, 2011, 7(3): 205-218.

[53] Xi X, Higgins D, Zechner K. A comparison of two scoring methods for an automated speech scoring system [J]. Language Testing, 2012, 29 (3): 371-394.

[54] Giang T L H, Kunnan A J. Automated essay evaluation for English language learners: A case study of MY access [J]. Language Assessment Quarterly, 2016, 13 (4): 359-376.

[55] Kang O, Johnson D. The roles of suprasegmental features in predicting English oral proficiency with an automated system [J]. Language Assessment Quarterly, 2018, 15 (2): 150-168.

(责任编辑:陈丽琼)