现代语言测试与评估丛书

曾用强 总主编

# 高校英语专业四级测试写作评分标准的设计与效度研究

李清华 著

科学出版社

现代语言测试与评估丛书

曾用强　总主编

# 高校英语专业四级测试写作评分标准的设计与效度研究

李清华　著



科学出版社
北京

**图书在版编目(CIP)数据**

高校英语专业四级测试写作评分标准的设计与效度研究 / 李清华著. —北京：科学出版社，2014.1
（现代语言测试与评估丛书 / 曾用强主编）
ISBN 978-7-03-039671-6

I. ①高… II. ①李… III. ①大学英语水平考试–写作–评分–标准–研究 IV. ①H315-42

中国版本图书馆 CIP 数据核字(2014)第 017487 号

责任编辑：刘彦慧　张翠霞 / 责任校对：郑金红
责任印制：钱玉芬 / 封面设计：无极书装

联系电话：010-6401 9074　电子邮箱：liuyanhui@mail.sciencep.com

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
双青印刷厂印刷
科学出版社发行　各地新华书店经销
*
2014 年 2 月第　一　版　开本：A5 (890×1240)
2014 年 2 月第一次印刷　印张：9 3/8
字数：300 000

**定价：58.00 元**

（如有印装质量问题，我社负责调换）

# 总　序

据 Spolsky 考证，正式的语言测试起源于中国东汉时期的科举考试。但是，现代意义上的语言测试却于 20 世纪中叶诞生在英美等发达国家。20 世纪 60 年代，外语测试作为一门新的学科从外语教学中独立出来。作为语言学、教育与心理测量学、计算机技术等的交叉学科，半个世纪以来，国外大批语言学家在语言测试领域取得了卓越的成就。两本专业期刊——《语言测试》(*Language Testing*)和《语言评估季刊》(*Language Assessment Quarterly*)相继诞生，在应用语言学界的影响越来越大，语言测试已经成为应用语言学领域的显学之一。在我国这样一个具有浓郁考试文化的大国，考试一直备受国人青睐。中国的高考和大学英语考试(CET)虽有数百万考生规模，但在研究方面仍落后于英美等发达国家。至今，以美国 ETS(Educational Testing Service)开发的 TOEFL 和英国剑桥大学考试委员会(CESOL)开发的 IELTS 为代表的西方国家的研究水平仍执全球之牛耳。但是，我们欣喜地看到，国内学者没有妄自菲薄，外语测试研究日益受到人们的重视，并已经和正在取得一些成绩，一些学者的论文在国际期刊发表，国内专业期刊《外语测试与教学》也在 2011 年于上海外国语大学问世。虽然有若干著述出版，但它们大多关注于测试的开发实践，而对测试理论与实践的研究较少。在此背景下，我和我的博士生们策划了“现代语言测试与评估丛书”。在国内学界享有盛誉的科学出版社高瞻远瞩，大力扶持外语研究，欣然同意出版这套丛书，可谓语言测试界的盛举。丛书为国内语言测试研究者提供一个平台，系统展示国内外语言测试领域的新成果，特别是国内学者的原创研究，供广大同行分享。

本丛书分为两个系列：①语言测试与评估：研究系列；②语言测试与评估：实践系列。前者以理论和研究为重点，主要面向应用语言学的研究生和语言测试研究者，后者以实践和应用为重点，主要读者是广大外语教师和教育行政管理人员。本丛书的第一系列围绕语言测试的热点和经典问题展开，主要涉及以下话题：语言测试的效度理论与实证研究、基于任务的语言测试研究、形成性评估研究、ESP 测试研究、动态评估研究、语言测试的评分研究、语言测试计算机自动评分研究、基于语料库的语言测试研究、语言测试的后效研究等。每一本书在保持各自特色的前提下，主要内容包括：系统介绍相关理论；本领域的主要研究方法；实证研究成果分析；附录：本领域近十年的研究成果。

本丛书的编著者均为语言测试方向的博士和知名学者，相信本丛书的出版将进一步促进国内语言测试研究的发展。本丛书是一个开放的平台，欢迎广大同仁提供自己的新作，欢迎广大读者提出批评与建议。本丛书编著者愿与国内同行一起，为使我国从考试人数大国早日发展成为考试研究大国而努力。

曾用强

2013 年于广州

# 前　言

高校英语专业四级测试(Test for English Majors，Band 4，TEM 4)是依据《高等学校英语专业英语教学大纲》和《高校英语专业四级考试大纲》设计实施的，考试的目的是全面检查已完成英语专业基础阶段课程的学生是否达到了教学大纲所规定的各项英语专业技能要求。这项考试自 1992 年实施以来，受到高校师生和社会的广泛认可，对外语教学和英语专业人才培养做出了巨大贡献。现行的英语专业四级写作测试的评分标准是基于当时的文献和教学大纲由专家设计的，从 1994 年实施以来，尚未进行系统的效度研究。该标准采用整体式评分法，而现有研究表明，整体式评分法不太适合 ESL/EFL 学习者(Bacha，2001；Shi，2001；Hamp-Lyons，1995；Hamp-Lyons，1991；Kroll，1990)。TEM 4 属于教学检查类测试，就应当为教学提供充分的反馈信息。在这方面，分项式评分法显然优于整体式评分法(Shaw & Weir，2007；Weigle，2002)。

本书报告了 TEM 4 写作评分研究的成果。基于语言测试效度整体观理论和写作测试效度验证的框架，提出本课题的理论框架。在全面综述二/外语写作能力理论研究、二/外语写作文本特征研究、二/外语写作评分研究等有关领域的研究成果的基础上，提出本研究的问题和研究方法。根据大规模问卷调查的结果和统计分析，作者设计了 TEM 4 分项式评分标准，然后，运用多种统计方法，特别是基于项目反应理论(item response theory，IRT)的多层面 Rasch 模型(multi-facet Rasch measurement)，并结合质的研究方法，对评分标准终稿进行效度验证。研究发现：总体而言，在评分员之间的一致性、评分员内部的一致性、评分员与受试之间的交互作用、评分量表的区分性等方面新标准均好于原标准，评分员

倾向于选择新的分项式评分标准。他们认为，新标准的优势主要是，能够为写作教学提供更详尽的反馈信息，指导和促进教学，能够全面反映学生的写作能力，而且根据新标准打分更加准确。比较新评分标准的维度和评分员的实际评分策略发现，评分标准中的所有要求都在评分实践中得到反映。

本研究提出的 TEM 4 的分项式评分标准在经过充分验证之后，已应用于 TEM 4 写作的评分实践中，在一定程度上促进了英语专业四级测试的发展。另外，国内大规模、高风险语言测试的写作部分的评分标准尚无基于实验证据设计并加以效度验证的研究，本研究亦为类似研究提供了范例。

本书是作者在上海外国语大学博士后流动站工作的部分成果。在博士后研究期间，导师邹申教授给予了作者悉心的指导和关心。上海外国语大学、广东外语外贸大学、上海大学、上海师范大学、南京国际关系学院等高校的数位专家对本研究提出了指导意见并参与了部分实验。来自全国的英语专业四级考试阅卷教师近 200 人参加了问卷调查，其中的 20 位教师还参加了阅卷实验。对此，作者表示衷心的感谢！本书得以与读者见面，依赖于“现代语言测试与评估丛书”编委会的资助，作者特别感谢作者的博士生导师、丛书总主编曾用强教授的支持和帮助！

科学出版社致力于扶持学术研究，刘彦慧女士为本书的编辑付出了巨大心血，作者深表谢忱！

作者还要感谢好友常州大学孔文教授和重庆师范大学彭康洲博士的鼎力帮助，感谢爱人孔慧女士和爱女李迪的理解与鼓励！

由于作者水平所限，书中难免存在疏漏与讹误，恳请广大同仁不吝批评指正。

李清华

2013 年于广州

# 目　录

# 表 目 录

# 图 目 录

第 1 章

# 引　言

高校英语专业四级测试(Test for English Majors，Band-4，TEM 4)是依据《高等学校英语专业英语教学大纲》和《高校英语专业四级考试大纲》设计实施的，考试的目的是全面检查已完成英语专业基础阶段课程的学生是否达到了教学大纲所规定的各项英语专业技能要求，考核学生综合运用各项基本技能的能力，以及学生对语法结构和词语用法的掌握程度。因此，本考试属于标准参照性、教学检查类测试。考试范围包括考试大纲所规定的听说读写技能及语法、词汇知识。这项考试自 1992 年实施以来，受到高校师生和社会的广泛认可，对外语教学和英语专业人才培养做出了巨大贡献。到 2012 年，参加考试的人数已达 27 万(邹申，2012)。作为一项大规模、高风险测试，其质量必须得到保证，而效度验证则是质量检验的必要环节。1993~1997 年的效度研究证明，1995 年版的四级测试具有较高的效度(邹申，1997)。随着《高等学校英语专业英语教学大纲(2000 年修订)》的颁布和《高校英语专业四级考试大纲(2004 年新版)》，测试内容和方法都有较大调整。对新版 TEM 4 的效度研究已经起步，例如，邹申(2005)、孔文(2009)、孔文和李清华(2009)等对 TEM 4 阅读部分的探索；彭康洲和李清华(2009)对 TEM 4 听力部分的分析。近十几年来，语言测试效度理论取得突破性发展。分类效度观正在逐渐被效度整体观取代(李清华，2006a；Chapelle，1999)。计算机技术的飞速发展也为教育与心理测量和语言测试提供了新的强大的研究工具。鉴

于 TEM 4 写作测试部分的效度验证还未进行，在新的效度验证框架下，对目前使用的英语专业四级写作测试展开系统的效度研究显得非常迫切。

任何一项测试，特别是大规模、高风险测试，都必须提供其效度的足够证据，其分数的意义才可能得到合理解释和正确使用。在新的效度验证框架下，效度是指根据测试分数对受试的语言能力做出的推断的有效性，效度的验证需要多方面证据的支持。

基于行为(performance-based)的语言测试(如写作测试、口语测试等)在提高效度的同时，由于其本身固有的复杂性和多层面性(multi-facetedness)，也难免引入一些误差因素(Schoonen，2005)。对于写作测试而言，对受试的写作能力的判断取决于评分员基于评分标准给出的分数。因此，如何制定评分标准使其反映写作能力的构念、评分员如何使用评分标准、评分员的内部及之间的一致性、写作测试任务的特征、受试作文文本的特征等成为写作测试效度研究的核心问题。在评分过程中，测试任务、评分方法、评分标准、评分员的评分策略、评分风格等方面的变量对评分员的评分决定起着很大作用(Weigle，2002)。国内外的语言测试界已经对该领域做了探索，并取得了一些成果。

现行的英语专业四级写作测试的评分标准是基于当时的文献和教学大纲由专家设计的，从 1994 年实施以来，尚未进行系统的效度研究。该标准采用整体式评分法，而现有研究表明，整体式评分法不太适合 ESL/EFL 学习者(Bacha，2001；Shi，2001；Kroll，1990；Hamp-Lyons，1995；Hamp-Lyons，1991)。TEM 4 既然属于教学检查类测试，就应当为教学提供充分的反馈信息。在这方面，分项式评分法显然优于整体式评分法(Shaw & Weir，2007；Weigle，2002)。在评分实践中，很多评分员发现这一评分标准不够具体，有些维度存有歧义。鉴于此，有必要依据 ESL/EFL 写作能力理论和其他著名 ESL/EFL 写作测试的评分标准为 TEM 4 写作测试设计新的分项式评分标准，并在当代语言测试效度验

证框架下对其效度加以验证，经过充分的验证之后，将这种评分标准应用于 TEM 4 写作的评分实践。因此，本书报告了 TEM4 写作测试分项式评分标准的设计与效度验证的研究成果。本书内容是 TEM 4 效度验证的一部分，可以在一定程度上促进英语专业四级测试的发展。另外，国内大规模、高风险语言测试的写作部分的评分标准尚无基于实验证据设计并加以效度验证的研究，本研究亦为类似研究提供范例。

第 2 章

# 二/外语写作测试效度验证的理论框架

## 2.1 语言测试效度整体观

效度(validity)和信度(reliability)是语言测试及其他教育与心理测量质量评价的根本要求。在语言测试领域，从普通计量学引入的信度概念比较稳定，教育与心理测量学家关注的是信度的估算方法；效度则是社会科学计量学的概念，自 20 世纪 30 年代提出以来，效度概念不断演变，在近半个多世纪取得了重大进展。尤其是从 20 世纪 80 年代中期以来逐渐形成的“效度整体观”(unified validity)颠覆了“分类效度观”的统治，把信度置于效度的框架之下，对当代效度研究产生了深远影响(Weir，2005；Messick，1989)。

在语言测试界，效度理论的演变可大致分为三个阶段。

(1)20 世纪 60~70 年代：效度三分观(trinitarian doctrine)。Lado(1961：321)对效度的经典定义是：“一项测试是否测量了它所要测量的东西？”效度和信度是语言测试的两大重要质量标准。信度就是稳定性(consistency)，是效度的前提条件。证明效度的最典型的方法就是“……在多大程度上这项测试与其他有效而又可靠的语言测试相关”(Oller，1979：417-418)。据 Angoff(1988)统计，在 20 世纪 30~80