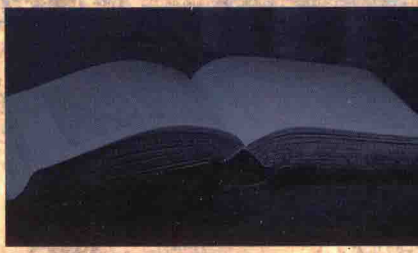
外教社
博学文库

外语教学测试构念研究

——以TEM-8作文评分员为例

Assessment Construct in Foreign Language Teaching:

The Case of Chinese Assessors of High-Stakes Exam Essays Written in English



陈建林 著

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

外教社
博学文库

教育部人文社会科学研究青年基金项目
（项目编号：15YJC740004）

外语教学测试构念研究

——以TEM-8作文评分员为例

Assessment Construct in Foreign Language Teaching:
The Case of Chinese Assessors of High-Stakes Exam
Essays Written in English

陈建林

上海外语教育出版社

图书在版编目（CIP）数据

外语教学测试构念研究：以TEM-8作文评分员为例/陈建林著.
—上海：上海外语教育出版社，2015
（外教社博学文库）
ISBN 978-7-5446-4079-4

Ⅰ.①外… Ⅱ.①陈… Ⅲ.①大学英语水平考试-写作-评分-研究
Ⅳ.①H315

中国版本图书馆CIP数据核字（2015）第217123号

出版发行：上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话：021-65425300（总机）
电子邮箱：bookinfo@sflep.com.cn
网　　址：http://www.sflep.com.cn http://www.sflep.com
责任编辑：许进兴

印　　刷：上海市崇明县裕安印刷厂
开　　本：890×1240　1/32　印张 10.375　字数 302 千字
版　　次：2015 年 12 月第 1 版　2015 年 12 月第 1次印刷

书　　号：ISBN 978-7-5446-4079-4 / G · 1302
定　　价：29.00 元

博学文库
编委会成员

（按姓氏笔画为序）

姓　名	学　校
王守仁	南京大学
王腊宝	苏州大学
王　蔷	北京师范大学
文秋芳	北京外国语大学
石　坚	四川大学
冯庆华	上海外国语大学
吕　俊	南京师范大学
庄智象	上海外国语大学
刘世生	清华大学
杨惠中	上海交通大学
何刚强	复旦大学
何兆熊	上海外国语大学
何莲珍	浙江大学
张绍杰	东北师范大学
陈建平	广东外语外贸大学
胡文仲	北京外国语大学
秦秀白	华南理工大学
贾玉新	哈尔滨工业大学
黄国文	中山大学
黄源深	上海对外贸易学院
程朝翔	北京大学
虞建华	上海外国语大学
潘文国	华东师范大学
戴炜栋	上海外国语大学

出版说明

上海外语教育出版社始终坚持“服务外语教育、传播先进文化、推广学术成果、促进人才培养”的经营理念，凭借自身的专业优势和创新精神，多年来已推出各类学术图书600余种，为中国的外语教学和研究作出了积极的贡献。

为展示学术研究的最新动态和成果，并为广大优秀的博士人才提供广阔的学术交流的平台，上海外语教育出版社隆重推出“外教社博学文库”。该文库遴选国内的优秀博士论文，遵循严格的“专家推荐、匿名评审、好中选优”的筛选流程，内容涵盖语言学、文学、翻译和教学法研究等各个领域。该文库为开放系列，理论创新性强、材料科学翔实、论述周密严谨、文字简洁流畅，其问世必将为国内外广大读者在相关的外语学习和研究领域提供又一宝贵的学术资源。

上海外语教育出版社

序

写作能力测评历来是语言测试的一个重要分支。在当前语言测试界愈来愈注重能力评价的大背景下，写作测试的重要性也日益凸显。然而，写作评分一直是困扰语言测试界的一大难题，由此引发了诸多有关写作评分的研究。该类研究无非涉及两个方面：分数和评分员。分数涉及测量结果的精确性或可靠性，而对评分员的研究则着重过程。陈建林的博士论文在对过程研究方面做了有意义的探索。

作者把作文评分员的评分过程视为一个问题解决过程。在文献回顾与先期实验的基础上，作者提出关于作文评分过程的假说：作文评分过程是在评分员先验知识与外界刺激相互作用中所建构的三个认知结构之间的互动。这三个认知结构分别是写作结构（writing structure）、评分结构（rating structure）和情景结构（contextual structure）。写作结构是评分员对不同能力水平作文应具备的特征的心理表征；评分结构是评分员在解决问题过程中所进行的一系列认知操作的心理表征；而情景结构是评分员对于社会环境因素所建构的心理表征。根据此假设所提出的评分过程模型，作者提出三个研究问题：第一，这三个认知结构是怎样的？第二，它们分别是如何建构的，影响建构的因素是什么？第三，作文评分过程的本质是什么？

作者以2012年英语专业八级考试（TEM-8）作文评分为例，从多个方面收集和分析数据，并得出以下主要结论：（1）影响作文评分过程的因

素主要有六个方面：评分标准、评分员培训、评分员特征、评分环境、文本特征和评分监督。上述六类因素可以进一步归纳为三个相互影响、相互作用的因素：评分机构因素、评分员特征与文本特征；(2) 数据分析验证了所提出的假设，即作文评分过程的本质是评分员建构的三个认知结构（写作结构、评分结构和情景结构）之间的互动；(3) 这三个认知结构的建构受到了诸多评分员个人和社会因素的影响。这些因素从不同方面、以不同程度对这三个结构的组成元素产生影响。作者的研究成果具有较高的理论意义和应用价值，有助于语言测试工作者了解评分过程的机制。

特别值得一提的是，作者有效地使用多种研究方法和分析手段，使结果更具有解释力度。比如利用多层面Rasch模型分析评分员的作文分数，采用SPSS和AMOS对问卷数据进行探索性和验证性因子分析，以及利用NVivo 8.0对有声思维和后续访谈录音进行转写、编码，并进行定量统计及定性分析。

本书理论阐述详尽，研究结果丰硕，是理论与实践相结合的典范。我深信，本书定会使语言教师和研究生读者受益匪浅。

邹　申
上海外国语大学教授、博导
教育部高等学校外语专业教学指导委员会委员
英语专业教学分指导委员会副主任委员
2014年11月22日

前言

目前，一些国际和国内的大规模语言考试中对于写作能力的测评均采用行为测试（performance assessment）方法，评分方式主要以人工评阅为主。那么，如何能够保证来自不同背景、具有不同经验和知识结构的评分员之间的评分一致性呢？要回答这个问题，首先需要对评分员的评分过程进行探究。

由于作者曾多次参加了英语专业四、八级考试的作文评分工作，在与同仁的交流过程中，对上述问题有了更加浓厚的兴趣。在2011年、2012年专业八级作文阅卷期间，作者通过深入访谈、有声思维、问卷调查等方式对几十位评分员的评分过程进行了调查，并收集他们评分的详细数据，在经过几年的整理、分析和探究后，终于形成此书。

本书以英语专业八级考试作文评分员为研究对象，对评分员的认知结构进行了描述，对评分过程的本质进行了理论概括，并就提高评分员评分一致性问题提出了看法和见解。本书共有七章。第一章对研究背景和研究意义做了陈述；第二章是文献回顾，主要对国内外有关评分员和评分过程的研究进行了梳理；第三章详细介绍了本研究的数据收集方法、步骤和分析过程；第四章和第五章是本书的核心部分。其中，第四章对评分员的评分过程进行了详细的描述，并归纳出指导评分员评分过程的三个认知结构，即写作结构（The Writing Structure）、评分结构（The Rating Structure）和情景结构（The Context Structure）。第五章对这三

个认知结构的建构及影响因素进行了概括，在此基础上提出了评分过程本质的框架模型；第六章阐述了本研究对语言测试理论和实践的启示；最后一章对本研究的结果和局限性做了总结。

本研究得到教育部人文社会科学研究青年基金项目“基于语料库的甘肃藏汉中学生英语书面语对比研究”（项目编号：15YJC740004）和“兰州大学中央高校基本科研业务费专项资金”（项目编号：2022014skzy001）资助。

本研究成果最终能成书，首先要感谢我的导师上海外国语大学邹申教授的悉心指导，也要感谢我在比利时鲁汶大学留学时指导老师Lies Sercu教授对本书的修改和建议。我还要感谢上海外国语大学许余龙教授、陈坚林教授、郑新民教授、张艳莉教授和上海理工大学刘芹教授的指导；感谢比利时鲁汶大学Kris Van Den Branden教授、Koen Jaspaert教授、Koen Van Gorp教授和英国兰卡斯特大学Tineke Brunfaut教授对本研究提出的宝贵意见；同时，我还要感谢中国农业大学赵瑛华博士、大连外国语大学泰中华博士、上海外国语大学姚涓涓博士、宁波大学丛迎旭教授、浙江师范大学郑连忠教授、上海外国语大学安德万同学，以及其他许多同仁和朋友的帮助与支持。我也要感谢上海外国语大学外事处李红玲教授、研究生部汪小玲教授、韩殿秀博士，以及张绍铎博士等在我申请上海外国语大学与比利时鲁汶大学联合培养双学位博士过程中的鼎力支持！

在这里也要特别感谢上海外语教育出版社将本书纳入“外教社博学文库”出版，并感谢匿名评审专家对本书提出的宝贵意见。

感谢我的妻子和家人的关心、支持和鼓励！

陈建林

2014年12月于兰州大学

List of Acronyms

AUA	Assessment Use Argument
CEFR	Common European Framework of Reference
CET	College English Test
CFA	Confirmatory Factor Analysis
CPE	Certificate of Proficiency in English
EFA	Exploratory Factor Analysis
EFL	English as a Foreign Language
EMT	Emergent Medical Technician
ESL	English as a Second Language
ESLP	English as Second Language Program
ESP	English for Specific Purposes
FCE	First Certificate in English
IELTS	International English Language Testing System
IRT	Item Response Theory
MCQ	Multiple-choice Questions
MELAB	Michigan English Language Assessment Battery
MFRM	Multi-faceted Rasch Model
NNS	Non-native Speaker
NS	Native Speaker
SLA	Second Language Acquisition
TAPs	Think-aloud Protocols
TBLA	Task-based Language Assessment
TEM	Test of English Majors
TOEFL	Test of English as a Foreign Language

Tables and Figures

Table of Contents

Chapter One

Introduction

1.1 Rationale for the study

Given the popularity of language testing practice and the importance of test results for testers and testees nowadays, it has often been wondered if international language testing services can guarantee that student essays written in a foreign language in very diverse learning contexts and rated and scored by many different locally-based raters are rated and scored in the same way. How can such tests, like IELTS or TOEFL, ensure that all test-takers are treated in exactly the same way? Do all raters base their rating on the same performance construct? Do they all assign scores with the same degree of lenience or strictness? It is exactly these questions that constitute the rationale for this study. In order to provide some background for answering these questions, it is necessary to begin with a brief review of the historical development of language testing practice itself.

Like the feature of the development of any entities, language assessment practice has also evolved from a pre-scientific stage when only the content of teaching was assessed. It was not until the 1960s that language assessment developed into a large-scaled, industrialized and