计算机化语言测试效度研究：基于证据的作文自动评分效度验证

Validation of Computerized Language Testing:
Evidence-based Validity Study of Automated Essay Scoring

高怀勇 著

復旦大學出版社
Fudan University Press

计算机化语言测试效度研究：基于证据的作文自动评分效度验证

Validation of Computerized Language Testing:
Evidence-based Validity Study of Automated Essay Scoring

高怀勇 著

复旦大学出版社
Fudan University Press

图书在版编目(CIP)数据

计算机化语言测试效度研究:基于证据的作文自动评分效度验证:英文/高怀勇著. —上海:复旦大学出版社,2018.8
ISBN 978-7-309-13770-5

Ⅰ. 计… Ⅱ. 高… Ⅲ. 计算机应用-英语-写作-评分-研究 Ⅳ. H315-39

中国版本图书馆 CIP 数据核字(2018)第 153781 号

计算机化语言测试效度研究:基于证据的作文自动评分效度验证
高怀勇 著
责任编辑/唐 敏

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编:200433
网址:fupnet@fudanpress.com http://www.fudanpress.com
门市零售:86-21-65642857 团体订购:86-21-65118853
外埠邮购:86-21-65109143 出版部申话:86-21-65642845
当纳利(上海)信息技术有限公司

开本 787×960 1/16 印张 22 字数
2018 年 8 月第 1 版第 1 次印刷

ISBN 978-7-309-13770-5/H·2831
定价:68.00 元

前　言

语言测试是语言教育、教学中至关重要的一环，是评估教学效果，提供教学反馈，进行教学决策的关键依据。语言测试与评价的质量取决于测试的效度（Aderson et al，1995；Bachman，1990；Messick，1989，Weir，2005；Bachman & Palmer，2016）。效度问题是整个测试工作的核心问题（修旭东，2010：11），是测试评价中最重要的考虑因素，是测试工作者要保证的最基本的测试质量特性（Bachman & Palmer，2016）。

在不断发展的计算机技术、语言学、认知科学和测量学的影响下，当代语言测试在试题管理、测试任务特征、受试的测试表现（test performance）、测试任务、测试后效、语言输入与输出等方面都发生了显著变化。特别是测试评分变化更为显著。产出性语言（如英文写作）的评分将逐渐（或已经开始）由计算机完成，传统测试与效度研究手段将面临巨大的挑战。面对新的计算机化测试手段，传统效度研究方法显得力不从心、捉襟见肘。计算机技术结合语言测试，特别是主观题目的测试（如作文自动评分），使得测试行为本身更加复杂化。通过计算机进行的作文自动评分会受到与构念无关因素（如，计算机配置差异、受试计算机技能差异、对计算机测试方式的焦虑程度等）的影响。随着我国大规模、高风险测试（如高考、国家公务员考试）的增加，测试在社会中扮演着越来越重要的角色。随着科技的发展，特别是计算机科学的发展，如何运用计算机来评估主观化语言测试项目已经逐渐成为现实。然而，到目前为止，计算机化主观测试题的评分原理与评分策略大多数是人工评分原理与策略的简单、机械模拟，或者是从主观测试题中抽取、提炼影响测试质量的指标作为计算机评分依据。这样的模拟与指标提取是否准确、有效，其合理性如何，能否代替人工评分？带着这些问题，该课题以风靡全国的句酷在线作文自动评分系统（JAESS）为研究对象，严格遵循理论密切联系实践的原则，以语言测试整体效度观为理论指导，深入研究了计算机化作文自动评分的准确性、有效性、合理性和社会影响。

本书分别从应试者的应试心理活动与测试表现、测试评分的准确度与其他效标之间的关联性、测试的区分度和聚合度以及测试的社会影响等视角探索了计算机化作文自动评分的效度，并提出相关对策与建议。

1. 计算机化语言测试情景下应试者的应试心理活动与测试表现之关系。要考查作文自动评分测试的效度就必须考查作文自动评分测试的应试者应试心理活动是否反映写作测试理论构念，即考查应试者在参加计算机化写作测试情景下的写作心理活动是否与理论上的写作心理过程一致，以及这些心理活动的运用（参与）度是否与写作质量产出有关。写作能力构念包含一系列预设的写作心理活动（过程），应试者在写作测试过程中是否运用了这些心理活动，如果运用了，运用的程度如何？这些心理活动的运用是否能反映在作文质量上，是否直接与产出作文质量有关？通过问卷调查、访谈、统计等一系列研究方法，该研究发现，应试者在参加计算机化的作文考试并自动评分的情况下，基本都运用了写作理论上预设的一系列心理活动，应试者的测试心理实际表现与写作理论预设心理活动基本一致。该研究进一步发现，虽然大多数应试者写作心理活动与作文产出质量正相关，然而，应试者的写作后检查（review）心理活动，特别是对“作文内容和词汇运用的适当性”的检查心理活动却与写作产出质量无关甚至负相关。此外，该研究发现，有少部分写作心理活动，如目标设定（goal setting）、主题与风格调整（topic & genre modifying）、写作材料组织（organizing），与应试者写作产出的质量低相关甚至无相关。以人的心理活动为研究对象难免会受到量表题目设置或者研究的操作过程甚至两者之间互动的影响。基于此，该研究提出建议认为：对于写作心理活动的考查，除了采用问卷、访谈等常规方法外，需要更加科学、有效的研究方法，如口头自陈（verbal report）、内省法（introspective）、回顾法（retrospective）等，才能更加准确地考查复杂的写作心理活动以及这些心理活动与写作质量产出之间的内在关系。尽管该研究发现，应试者的写作心理活动不能完美反映写作质量产出，但是写作任务能够诱导、激发所有的写作心理活动，大多数心理活动投入能够表现为写作产出的质量。基于此，该研究认为计算机化作文自动评分是建立在写作理论基础上的，具有充足的写作理论构念证据。

2. 计算机化语言测试评分的准确度与其他效标之间的关联性。要考查基于计算机的作文自动评分测试效度，该研究认为必须要调查该测试与其他相关效标之间的关系，如作文自动评分与相关人工评分之间的相关性、与其他标准测试作文能力评估的一致性。通过调查 JAESS 的作文自动评分与一系列非测试指标（如教师对应试者写作能力总体评价，应试者本人对自己的写作能力评价）和测试指标（如对应试者作文的人工评分）之间的关系，该研究发现，基于计算机的作文自动评分系统 JAESS 具有很高的评分效度：JAESS 评分与人工评分的一致性达到 75.8%，与教师对应试者写作能力总体评价的一致性达到了 87.3%，与应试者本人对自己的写作能力评价的一致性达到 75.8%，具有很高的评分准确率。基于此，该研究认为，以 JAESS 为代表的计算机化作文自动评分系统评分基本准确、有

效、合理。然而,在问卷和访谈中,该研究发现,应试者认为该评分系统有很大的优越性的同时也有不足的地方。优点在于该系统能准确有效评估词汇拼写、词汇固定搭配与作文语法错误;缺点在于该系统在诊断作文的谋篇布局、句子结构、汉式句子、文章薄弱点等方面并提出修改意见有需要改进的地方。针对这一发现,该研究提出了适当增加计算机化作文评分指标的提取与分类等解决问题的对策,并建议测试开发者和使用者针对本研究发现的缺点适当作出有针对性的改进与提高,从而提高计算机化作文自动评分的有效性和合理性,以提高作文自动评分后信息反馈的充分性和有用性。

3. 计算机化语言测试的区分度和聚合度。一项高效度的写作能力测试,不仅应该有聚合效度,更应该有区分效度,前者考查该测试与其他非考试写作能力指标之间的关系,后者考查该测试与其他非考试非写作能力指标之间的关系。通过考查基于计算机的作文自动评分系统 JAESS 的评分结果与一系列非测试写作能力指标(如教师对应试者写作能力总体评价,应试者本人对自己的写作能力评价)和非测试非写作能力指标(如应试者的非语言能力指标,如物理、化学、体育等)之间的关系,该研究发现,作文自动评分系统 JAESS 的评分结果与一系列非测试写作能力指标之间有很高的拟合度(与教师对应试者写作能力的总体评价为 87.3%,与应试者本人对自己的写作能力评价为 75.8%),即表明以 JAESS 为代表的计算机化作文自动评分系统具有很高的聚合度或者拟合度。与此同时,该研究也发现,作文自动评分系统 JAESS 的评分结果与一系列非测试非写作能力指标之间既没有正相关也没有负相关,或者相关在统计学上没有显著性意义。基于以上发现,该研究认为基于计算机的作文自动评分系统 JAESS 具有很高的聚合效度与区分效度,能够准确区分语言能力和非语言能力,对语言能力的检测具有敏感性。

4. 计算机化语言测试的社会影响。一项语言测试的测试后效(影响)在该测试的效度验证中起着举足轻重的作用。在此,测试的社会影响主要包括:1)测试的公平性,即该测试是否公平,是否对某些应试者有偏见,而对其他应试者有利;2)测试对教学的反拨作用,即测试对教学的微观影响,这些影响是积极的还是消极的,积极到什么程度,消极到什么程度等一系列问题;3)测试对社会的宏观影响,包括测试及其使用对教学机构、测试使用者和测试有关人员的影响。该课题通过问卷调查与小组访谈等研究方法调查了教学机构(如教育行政部门)和测试使用者(如应试者和教师)对计算机化语言测试的社会影响。研究表明:1)以 JAESS 为代表的计算机化语言测试具有测试的公平性,该测试没有因为"考生对计算机的熟悉程度""计算机水平"和"测试方式"等问题造成测试的不公和对部分考生的偏见;2)以 JAESS 为代表的计算机化作文自动评分语言测试系统对英语

写作教学主要表现为正面、积极影响。其正面反拨效应主要体现在该系统“能循序渐进逐渐提高学生写作能力”、“培养学生的写作兴趣”以及“提高写作教师的教学工作效率”等方面。此外,该系统的积极反拨作用还体现在该系统自动评分后所提供的信息反馈在许多方面有积极意义,对考生的学和教师的教有积极参考意义。如该系统在“词汇拼写错误检查”“词块使用”“词汇搭配”和“语法错误检测”等方面具有高效、准确、灵敏等特点。然而,该研究调查发现以 JAESS 为代表的计算机化作文自动评分语言测试系统有许多需要改进的地方。主要体现在该系统对作文的篇章结构(text organization)、内容、长句分析和汉式英语的检测不够准确和灵敏,尤其是在作文的篇章结构和英语长句分析上常常出现误差;3)在宏观层面上,以 JAESS 为代表的计算机化作文自动评分语言测试系统对社会的影响主要体现为积极、正面影响,虽然该系统尚有一定的缺陷,如对教学机构、社会群体的社会影响力不大。基于此,该课题认为以 JAESS 为代表的计算机化作文自动评分语言测试系统具有测试公平性,对教学有积极的反拨作用,对社会的影响主要体现为积极、正面影响,因此具有积极的测试后效。针对计算机化作文自动评分系统的社会影响,该研究提出建议:测试研究开发者针对系统的缺点进行改进,特别应该提高计算机化作文自动评分语言测试系统在检测、诊断作文篇章结构、内容和英语长句分析等方面的灵敏度和准确有效度,从而进一步提高其对教学的正面反拨效应和社会影响力,为计算机化作文自动评分语言测试系统在全社会的推广和普及做出有益探索和实践。

本研究的主要目的在于抛砖引玉,为计算机化语言测试在我国的推广和普及作出有益探索。该成果一方面能为众多计算机化作文自动评分研究增加新的例证,对提高计算机化作文自动评分的科学性和有效性作出有益探索和努力。另一方面,该成果对计算机化作文自动评分效度的探索和研究对计算机化作文自动评分的推广、应用以及语言测试效度理论建设都有一定的意义,有利于推动学术研究,促进现代化历史进程。该成果的价值在于对大规模高风险语言测试的计算机化评分的设计、开发、推广提供了有益的探索和建设性对策与建议

本书是作者国家社科基金项目的部分成果,在研究期间,来自陕西(陕西师范大学)、四川(四川农业大学、四川大学和西南交通大学)和上海(上海交通大学)的 60 多位英语老师和 500 多位非英语专业学生参加了问卷调查,其中有 20 位教师还参加了阅卷实验和访谈。对此,作者表示衷心的感谢!作者特别感谢上海交通大学博士生导师金艳教授的无私赐教,感谢导师西南交通大学金桂林教授的长期支持与帮助!

本书得以与读者见面,依赖于复旦大学出版社的大力支持。本书的编辑为此付出了巨大的心血,作者深表谢忱!

作者还要感谢好友西安外国语大学刘锋博士的鼎力帮助，感谢爱人王晓燕女士的理解与鼓励！

由于作者水平有限，书中难免存在疏漏与讹误，恳请广大同仁不吝批评指正。

高怀勇

2018 年 3 月

List of Abbreviations

AERA	American Educational Research Association
AES	Automated Essay Scoring
AI	Artificial Intelligence
APA	American Psychological Association
ALTE	Association of Language Testers in Europe
ANOVA	Analysis of Variance
BETSY	Bayesian Essay Test Scoring System
BIEAESS	Bingo Intelligent English Automated Assay Scoring System
BTC	Bayesian Text Categorization
CA	Correlational Analysis
CADLT	Computer Adaptive Language Testing
CALT	Computer Assisted Language Testing
CAMLA	Cambridge Michigan Language Assessment
CBLT	Computer-based Language Testing
CCT	Cloud Computing Technology
CET	College English Test
CET-4	College English Test Band 4
CET-6	College English Test Band 6
CLT	Computerized Language Testing
C-Rater	Conceptual Rater
CT	Computed Tomography
EEE	EFL Essay Evaluator
EFL	English as Foreign Language
E-rater	Electronic Essay Rater
ESOL	English for Speakers of Other Languages
ETS	Educational Testing Service
FA	Factor Analysis

FLTRP	Foreign Language Teaching and Research Press
GLM	General Linear Model
GMAT	Graduate Management Admission Test
GMAT AWA	Graduate Management Admissions Test Analytical Writing Assessment
GRE	Graduate Record Examination
HR	Human Rater
HS	Human Score
IEA	Intelligent Essay Assessor
IELTS	International English Language testing System
IEMS	Intelligent Essay Marking Systems
ILT	Internet-based Language Testing
JAESS	Juku Automated Essay Scoring System
KMO	Kaiser-Meyer-Olkin (test to assess the appropriateness of using factor analysis)
LSA	Latent Semantics Analysis
LSD	Latent Semantic Dimensions
MBM	Multivariate Bernoulli Model
MM	Multi-nominal Model
NAEP	National Assessment of Educational Progress
NCME	National Council on Measurement in Education
NLP	Natural Language Process
NNSF	National Natural Science Fund
NP	Ngee Ann Polytechnic
PEG	Project Essay grader
PETS	Public English Testing System
PKT	Pearson Knowledge Analysis Technology
PS-ME	Paperless School Free-text Marking Engine
QCES	Questionnaire about Consequential Evidence of JASS for Students
QCET	Questionnaire about Consequential Evidence of JAESS for Teachers
OICE	Outline of Interview for Consequential Evidence
OIWP	Outline of Interview for Writing Process
QNA	Questionnaire about Non-writing Ability
QWP	Questionnaire about Writing Process
RA	Regression Analysis

REA	Reliability Analysis
RMA	Rasch Model Analysis
SEAR	Schema Extract Analyse and Report
SCAU	Sichuan Agricultural University
SCU	Sichuan University
SEM	Structural Equation Modeling
SJTU	Shanghai Jiaotong University
SNU	Shanxi Normal University
SSEWT	Students' Self Evaluation of the Writing Task
SWJTU	Southwest Jiaotong University
TEOWA	Teachers' Evaluation of students' Overall Writing Ability
TEM	Test for English Majors
TLUD	Target Language Use Domain
TOEFL CBT	Test of English as a Foreign Language Computer-based Test,
TOEFL iBT	Test of English as a Foreign Language Internet-based Test
TOEIC	Test of English for International Communication
VSM	Vector Space Model
WISLT	Web-based Individualized Self-adaptive Language Testing

Contents

Appendix A－M
（p332－487）

List of Tables