

中青年学者外国语言文学学术前沿研究丛书

自动作文评阅系统 信效度及辅助教学研究

张荔 盛越 著

EXPLORING the *Reliability and Validity* of
Automated Writing Evaluation System and
Its Effect on English *Writing Instruction*

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

中青年学者外国语言文学学术前沿研究丛书

自动作文评阅系统 信效度及辅助教学研究

张荔 盛越 著

EXPLORING the *Reliability and Validity* of
Automated Writing Evaluation System and
Its Effect on English *Writing Instruction*

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

北京 BEIJING

图书在版编目(CIP)数据

自动作文评阅系统信效度及辅助教学研究: 汉、英 / 张荔, 盛越著. — 北京: 外语教学与研究出版社, 2018.4

(中青年学者外国语言文学学术前沿研究丛书)

ISBN 978-7-5135-9992-4

I. ①自… II. ①张… ②盛… III. ①英语—写作—教学研究 IV. ①H319.36

中国版本图书馆 CIP 数据核字 (2018) 第 093433 号

出 版 人 徐建忠

责任编辑 李婉婧

封面设计 高 蕾

出版发行 外语教学与研究出版社

社 址 北京市西三环北路 19 号 (100089)

网 址 <http://www.fltrp.com>

印 刷 北京九州迅驰传媒文化有限公司

开 本 650×980 1/16

印 张 15.75

版 次 2018 年 5 月第 1 版 2018 年 5 月第 1 次印刷

书 号 ISBN 978-7-5135-9992-4

定 价 56.90 元

购书咨询: (010) 88819926 电子邮箱: club@fltrp.com

外研书店: <https://waiyants.tmall.com>

凡印刷、装订质量问题, 请联系我社印制部

联系电话: (010) 61207896 电子邮箱: zhijian@fltrp.com

凡侵权、盗版书籍线索, 请联系我社法律事务部

举报电话: (010) 88817519 电子邮箱: banquan@fltrp.com

法律顾问: 立方律师事务所 刘旭东律师

中咨律师事务所 殷 斌律师

物料号: 299920001

本书为国家社会科学基金项目“基于语料库和云技术的网络自动作文评阅系统信效度及其辅助教学研究（13BYY081）”的成果。

丛书序

进入新世纪以来,我国外语学科的学术研究呈现繁荣的局面,不仅学界专家、尤其是中青年学者的研究热情持续高涨,而且申报项目、出版论著或者发表论文的数量和质量也大幅度提高。以全国哲学社会科学规划办公室提供的数据为例,2014年国家社科基金年度项目和青年项目申报总数为28,186项,其中30-45岁这一年龄段共有17,729人申报,占总申报量的62.9%;经过评审,共有2395人立项,占总立项数的62.7%。这表明包括外语学科在内的中青年学者已成为学术研究的中坚力量。与此同时,外语学科的研究者除了在国内外重要学术刊物发表高水平的论文之外,还不断开拓论著发表渠道,通过创办学术集刊等形式表达和传播学术思想。据统计,近年来,各高校创办的外语集刊达几十种之多,在国外刊物上发表论文的数量也逐年增加。形成这种局面的原因多种多样,但主要归功于研究者们探索新知的内在需求和国家激励学术创新的外部环境。

然而,由于种种原因,中青年学者仍然面临着专著出版难的问题。尽管他们大多在国内外接受过系统的学术训练,思想活跃、心智敏锐,能够迅速抓住学术前沿话题,撰写的专著具有较高的学术价值,但常常陷入出版难的窘境,即使几经周折得以出版,也是“零散型”的著作,淹没在书海中而难以产生较大影响,更无法形成“集约型”效应。从某种意义上来说,这是极大的智力资源浪费,也在一定程度上挫伤了中青年学者学术研究的积极性。

中青年学者是中国学术发展的希望。为了保护中青年学者的学术热情,推动他们在学术前沿勇于探索,多出成果,外语教学与研究出版社秉承“开放性、学术性、原创性、前沿性”的原则,设立“中青年学者外国语言文学学术前沿研究丛书”出版项目。“开放性”是指对全国各高等学校和科研院所开放,凡是希望通过这一项目出版外国语言文学类各领域研究成果的中青年学者,皆可申请加入;“学术性”是指研究成果具有深刻的学术思想和观点,能够反映外国语言文学各领域理论和实践的本质和规律;“原创性”是指发前人所未发、想前人所未想,在研究内容和形式上有创见和创新;“前沿性”是指研究成果反映外国语言文学各领域的最新发展动态,具有前瞻性。近年来,随着计算机技术的广泛应用

和全球化进程的进一步加快,外国语言文学学术研究的内容和形式发生了变化,并呈现交叉融合的趋势。因此,我们鼓励中青年学者以问题为导向,运用新技术、新方法和新视角,对语言学、文学、翻译学、外语教学等诸多方面进行跨学科研究,力争推出一批集学术性、原创性和前瞻性为一体的最新研究成果,构建具有中国特色的外国语言文学学术话语体系。

首批入选“中青年学者外国语言文学学术前沿研究丛书”的专著共有十本:《教育语言学——一个社会符号的模式》、《文学交际中的读者:叙事虚构作品解读的自由与局限》、《翻译诗学》、《中国晚期二语习得者英语被动句加工的认知神经机制:行为与脑科学的整合研究》、《萧乾文学翻译思想研究》、《詹姆斯·马丁访谈录》、《英汉语篇跨文化修辞研究》、《莎士比亚戏剧中强调语的语用分析》、《基于语料库的中国英语学习者强化语型式和意义研究》和《国家形象与外宣翻译策略研究》,涵盖语言学、文学、翻译、外语教学等不同领域。

本中青年学者外国语言文学学术前沿研究丛书注重质量和创新。首批入选的专著均经过专家审读和评阅。为确保丛书的学术品质,外语教学与研究出版社还将聘请资深专家担任系列丛书的顾问,并成立由外国语言文学学科领域知名学者组成的编委会,对申请出版的学术专著进行评议和遴选。本丛书项目将按照相关质量标准,组织专家对申报项目进行评审,入选者列入出版计划。

我们秉承严谨治学的态度,希望通过学术出版为学术传承与创新提供平台,但由于水平有限,书中恐有疏漏之处,恳请诸位专家和学者不吝指教。

外语教学与研究出版社
高等英语教育出版社
2015年9月

前言

随着语料库、云技术、自然语言处理、机器学习、人工智能技术的不断发展,现代技术对语言教学的作用日趋明显,其中一个重要的应用就是自动作文评分(AES)。国外比较常用的AES系统有PEG, IEA, LSA, E-rater, BETSY, IEA, Criterion, IntelliMetric等,对于AES的研究主要集中在三个方面:(1) AES的开发性研究;(2) AES的评分和评价功能信效度研究;(3) AES对写作教学的辅助作用研究。然而国外的AES是针对英语为母语的写作者,未必适用于二语写作者。国内AES的研究起步较晚,目前开发的系统主要有批改网系统、冰果英语智能作文系统与大学英语四、六级作文自动评分系统等。然而针对我国AES系统的可靠性评定及其对教学辅助作用的系统性研究几乎没有,因此本书呈现了对我国自主研发的基于语料库和云技术的网络自主作文评阅系统——批改网系统(<http://pigai.org/>)的信效度及其辅助教学研究,以填补这方面的不足。

本书共十二章,分理论和评述性研究、信效度研究和辅助教学研究三个方面,具体为:(一)自然语言处理与语言学关系的理论性研究和国外多种AES系统的评述性研究;(二)国内自主研发的AES系统批改网的信效度个案研究、基于CLEC语料库作文的信度研究以及批改网测量性、归纳性和外推性信效度研究;(三)批改网使用情况的问卷调查研究、不同水平学习者使用效果的差异性研究、作文修改过程的文本分析研究、三种反馈比较研究以及包含AES的学术英语写作课程形成性评估模式的研究(其中部分内容已在学术期刊公开发表)。

本书采用CET考试作文样卷进行个案分析,利用CLEC语料库作文进行大规模研究,从点到面,以不同视角对系统信度进行检验,又从测量性、归纳性、外推性、解释性、衍生性和实用性等对批改网加以信效度分析,在内容和方法上具有一定的特色和创新性。本研究还首次在全国范围内从任课教师和不同年龄层次学生的角度实施了对写作教学和反馈的现状和需求、对使用计算机评阅的感受和评价以及批改网系统使用的问题和改进建议的大规模问卷调查。此外,研究从准确性、流利度、复杂性和作文总体质量四个方面探究了系统对不同水平学生英语写作能力的影响;就学生在批改网平台上作文修改过程和修改效果开展研究,以

洞察学生如何使用批改网对作文进行修改的细节过程；针对二语写作中的教师、同伴、AES反馈的比较研究侧重于研究学生的感知及采纳度，从一定程度上填补了现有研究的不足；将AES反馈运用于学术英语写作教学，将同伴反馈和教师反馈相结合的形成性评估模式体现了AES反馈与其他反馈形式相结合的可能性。

本研究为国家社科基金项目“基于语料库和云技术的网络自动作文评阅系统信效度及其辅助教学研究”的成果，本人在此表示感谢。同时也要感谢美国加州大学的Mark Warschauer教授、美国明尼苏达大学的Richard Beach教授、香港理工大学的李蓝教授在合作研究中所给予的关注和支持；感谢上海交通大学外国语学院的支持和帮助以及黄绍莹、陈冰青、陈心怡、陈德凤等几位研究生的帮助或参与；感谢上海交通大学教育发展中心对研究项目的资助和研究过程的协助；特别感谢批改网系统工作人员所给予的大力支持；最后还要感谢家人的默默付出——正是有了所有人的支持，才使得项目的完成和专著的出版成为可能。

目 录

第一章 引言	1
1.1 研究目的	1
1.2 主要内容	3
第二章 自然语言处理与语言学理论及方法论的历时和共时分析	9
2.1 引言	9
2.2 自然语言处理的定义	10
2.3 自然语言处理与语言学理论关系的历时分析	10
2.4 自然语言处理与语言各个层面的共时分析	16
2.5 方法论	18
2.6 结语	21
第三章 AES系统研究述评与展望	23
3.1 引言	23
3.2 AES系统综述	23
3.3 AES系统比较与展望	34
3.4 结语	38
第四章 AES系统反馈效果个案研究	39
4.1 引言	39
4.2 理论及相关研究背景	39
4.3 研究过程和方法	42
4.4 研究结果	42
4.5 讨论	45
4.6 结语	49

第五章 批改网AES系统信度评估 ——基于CLEC语料库的研究	51
5.1 引言	51
5.2 AES系统信度研究综述	51
5.3 研究过程和方法	53
5.4 研究结果	54
5.5 讨论	56
5.6 结语	57
第六章 AES系统测量性、归纳性和外推性效度研究	61
6.1 引言	61
6.2 研究背景	61
6.3 研究方法	65
6.4 研究结果	66
6.5 讨论	69
6.6 结语	71
第七章 AES系统使用现状调查研究	73
7.1 引言	73
7.2 文献综述	73
7.3 研究方法	76
7.4 研究结果	79
7.5 讨论	99
7.6 结语	103
第八章 AES系统对不同水平学习者英语写作能力的影响	105
8.1 引言	105
8.2 文献综述	106
8.3 研究设计	108
8.4 研究结果	110
8.5 讨论	113
8.6 结语	115

第九章 基于AES系统反馈的修改过程研究	117
9.1 引言	117
9.2 文献综述	118
9.3 研究方法	121
9.4 研究结果	123
9.5 讨论	131
9.6 结语	135
第十章 中国学生对教师、同伴、AES反馈的感知及 采纳度研究	137
10.1 研究背景	137
10.2 理论框架	138
10.3 文献综述	142
10.4 研究方法	144
10.5 研究结果	146
10.6 讨论	153
10.7 结语	157
第十一章 融入AES的学术英语交际课程形成性评估模式 和效果研究	159
11.1 引言	159
11.2 文献综述	160
11.3 研究方法	163
11.4 研究结果	165
11.5 讨论	170
11.6 结语	172
第十二章 结语	173
12.1 研究内容总结	173
12.2 研究意义	176
12.3 存在不足和进一步研究方向	178
参考文献	181
附录	207

图目录

图3.1	PEG评分方框图	25
图3.2	基于三个部分个性化结合及辅助措施的IEA评分示意图	26
图3.3	IntelliMetric评分要素模型	28
图3.4	IntelliMetric结构图	28
图3.5	E-rater评分要素分解图	30
图3.6	C-rater评分步骤	34
图3.7	基于认知的写作能力评估模型	37
图7.1	班级规模	79
图7.2	布置作文频率	79
图7.3	写作重要性	80
图7.4	强化训练作用	80
图7.5	作文批改占据时间	80
图7.6	学生写作频率	82
图7.7	强化训练作用	82
图7.8	写作重要性	82
图7.9	及时反馈	82
图7.10	喜爱的反馈种类	83
图7.11	希望的反馈方式	83
图7.12	师生对批改网具体反馈评价比较	89
图8.1	批改网整体反馈界面	106
图8.2	批改网按句点评	107
图8.3	高低水平组学生对反馈方式呈现的选择	112

图10.1 学生最喜爱的反馈类型	147
图10.2 学生认为最有效的反馈类型	148
图10.3 学生对三种反馈有效性感知的频率	148
图10.4 语言、内容、结构方面三种反馈的频率	150
图11.1 学术英语交际课程形成性评估模式及研究设计	163

表目录

表1.1	AES效度框架.....	4
表3.1	五种AES系统的个性和共性比较.....	35
表4.1	样卷人/机评阅的一致性检验.....	42
表4.2	样卷人/机评阅的相关性检验.....	43
表4.3	系统对作文的总体评价.....	43
表4.4	系统对作文的分类评价.....	44
表4.5	两组样卷高低分作文系统反馈类型的频数统计.....	45
表5.1	语料库人工评阅与系统评阅的一致性检验.....	55
表5.2	语料库人工评阅,重新人工评阅与系统评阅的相关性.....	55
表5.3	语料库人工评阅,重新人工评阅与系统评阅的平均分 与标准差.....	56
表6.1	自动评分系统与人工评分的描述统计.....	67
表6.2	三次作文人工评分与系统评分的相关性和一致性.....	67
表6.3	系统三次作文评分的相关性和人工三次作文评分的 相关性比较.....	68
表6.4	系统三次作文评分的差异性(One-way ANOVA).....	68
表6.5	系统三次作文评分差异性两两比较(Post Hoc).....	68
表6.6	人工作文评分与听说读及档案袋得分的相关性.....	68
表7.1	参与教师基本情况描述.....	76
表7.2	参与学生基本情况描述.....	76
表7.3	教师调查问卷信度.....	78
表7.4	学生调查问卷信度.....	78
表7.5	人工与计算机反馈情况.....	81
表7.6	教师使用批改网总体感受.....	83
表7.7	教师对批改网具体反馈的评价.....	84
表7.8	学生使用批改网总体感受.....	86
表7.9	学生使用批改网修改情况.....	87
表7.10	学生使用批改网各分项评价.....	88

表7.11 不同层次学生对问卷调查项目认同情况的差异性	90
表7.12 不同层次学生对问卷调查项目感知差异性两两比较 (Post Hoc)	94
表8.1 低水平组前后测作文数据的配对样本t检验结果	110
表8.2 高水平组前后测作文数据的配对样本t检验结果	111
表8.3 高低水平组学生对批改网的态度	111
表8.4 高低水平组学生对不同反馈方式的喜爱程度	112
表9.1 反馈类型	123
表9.2 各类反馈频数	126
表9.3 细节反馈排名	126
表9.4 修改次数	127
表9.5 修改次数人数统计	128
表9.6 初稿和终稿长度成对样本统计量	128
表9.7 修改类型和修改层面频次及修改示例	128
表9.8 修改层面和修改类型交叉表	131
表9.9 初稿和终稿成对样本检验	131
表10.1 李克特量表的信度分析	146
表10.2 学生对不同反馈类型喜爱度的频数和频率	147
表10.3 学生对只收到某种反馈的接受度	149
表10.4 学生对完全采纳某种反馈的接受度	149
表10.5 三种反馈正面反馈和负面反馈的频数和频率	150
表10.6 三种反馈语言、内容、结构反馈的频数和频率	150
表10.7 学生对三种反馈的采纳情况	152
表10.8 学生对语言、内容和结构反馈的采纳情况	153
表11.1 短期、中期和长期形成性评估对教学的调整	166
表11.2 实验班和对照班调查问卷各项认同度频数和差异性检验	168
表11.3 实验班论文初、二、终稿总分和各维度分项分重复测量 方差分析	169
表11.4 论文初、二、终稿成对比较	169
表11.5 实验班和对照班论文终稿总分和各维度分项分差异性比较	170

第一章 引言

1.1 研究目的

随着人工智能的进一步发展,自然语言处理技术不断提升,其应用范围也更加广泛。其中一个针对语言写作教学的应用就是自动作文评分(Automated Essay Scoring, AES)或自动写作评阅(Automated Writing Evaluation, AWE)。AES系统的研究始于20世纪60年代的美国,比较常用的系统有PEG、IEA、E-rater、IntelliMetric、BETSY等。其研究主要集中在以下三个方面:(1) AES的开发研究;(2) AES的评分和评价功能信度和效度研究;(3) AES对写作教学的辅助作用研究。

国外有关AES的开发研究集中在20世纪,后来的研究多数为对现有系统的验证和完善以及系统对写作教学辅助作用的研究。研究者对AES往往有比较积极的看法:例如Ramineni和Williamson(2013)发现E-rater对大规模写作评分有积极作用,可辅助人工评分;Attali等(2012)认为使用AES不仅能增加人工评阅的可靠性,而且人工评阅与机器评阅的相似性也越发接近;Foltz等(1999)验证了AES与人工评阅一样准确;甚至人机评阅的相关性大大高于人工评阅者之间的相关性(Folzeet al., 2013)。AES对写作教学的辅助作用研究是在AES可靠性和有效性前提下对该领域研究的进一步深入:Burstein等(2004)分析了Criterion对作文评分和评介两方面作用,指出该系统能帮助学生提高写作技能,但需在教师指导下运用;Klobucar等(2013)认为Criterion能有效帮助差生提高写作水平;Lee等(2011)发现LSA可以从内容和结构上给学生即时反馈,帮助成人学生提高写作质量。当然也

有一些研究者对AES有所质疑：Condon (2013) 对AES结构效度提出质疑，认为AES测量的不是真实意义上的写作；McCurry (2010) 认为AES只适合于限定的写作任务，对于更加宽泛的写作任务不如人工评分可靠性高。但总体上来说，AES的可靠性、有效性及其对教学的作用得到了许多学者的肯定，因此像E-rater这种AES系统就被使用于GRE、GMAT、TOFEL等高风险大规模考试的作文评分中 (Dikli, 2006)，或与人工评阅分数加以平均获得最终分数，或作为人工评阅的质量检测方法 (Bridgeman, Trapani & Attali, 2012)。

然而国外的AES针对英语为母语的写作者，未必适用于二语写作者。国内AES的研究起步较晚：梁茂成 (2010, 2011) 构建了中国学生英语作文自动评分模型，研制了大规模考试英语作文自动评分系统；贺俊杰等 (2012) 采用因子分析和多元回归分析法构建了面向中国英语学写者的写作计算机辅助评分模型。而批改网 (www.pigai.org) 这一我国自主研发的AES系统在很多大中学校得到越来越多的运用，相应的研究层出不穷。然而系统一经开发和运用，必须经过严格的检验，以帮助人们揭示系统是否能真正有效地测量写作能力 (Kane, 2013)。但国内对于AES的评分和评价功能信度和效度研究比较少，仅杨婷婷 (2012) 通过两种国内开发的AES进行评分相关性分析后认为这两种AES评分不一致。而AES对写作教学的辅助作用研究相对较多，例如石晓玲 (2012) 以批改网为例，分析了在线AES对写作教学的作用，并提出应增强在线AES的智能性及构建多元写作反馈机制；顾成华和王丽 (2012) 认为AES有助于提高学生的写作水平，但只是辅助教学的工具，需不断完善。这些研究往往角度比较单一，没有针对基于批改网反馈的作文修改过程的文本分析，也没有对批改网使用状况与写作教学的模式进行深入探讨。

从上述分析来看，国内现有的AES系统一批改网的信效度和辅助教学的系统性研究目前还没有。相关的可靠性评定比较缺乏，系统是否能真正有效实现作文自动评阅一直备受关注和质疑，急需对其进行信度和效度的研究。而AES系统信度和效度的研究是其辅助写作教学的前提 (Warschauer & Ware, 2006)，因此本研究对批改网实施信效度检验，不仅是对测试领域信度和效度理论的进一步完善，也有利于后续辅助教学的研究。此外，从系统辅助教学研究角度来看，对基于AES系统的作文修改过程的文本研究有利于动态了解系统如何辅助教学，有何不足。对三种反馈方式的比较研究有利于进一步了解学生对各种反馈方式的接受度