Modern Linguistics 现代语言学, 2026, 14(2), 13-20
Published Online February 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.142108

Hans汉斯

英语写作自动评分系统研究概述

汪鑫济

上海大学外国语学院，上海

收稿日期：2025年12月24日；录用日期：2026年1月16日；发布日期：2026年1月28日

摘　要

由于大规模考试数量的增多和考试人数的增加，传统的人工阅卷不再能够满足跨地域、跨学科的大规模考试阅卷，各地人工阅卷工作不仅给评阅教师增加了极大的负担，同时也大大降低了评阅工作的效率。随着自然语言处理技术的不断成熟，自动评分系统有望成为解决这一难题的希望。作为现代教育技术的自动评分系统在云端技术的支持下，以语料库为基础，对学生的答卷进行评分和反馈，这不仅符合了“重视现代信息技术应用，丰富英语课程学习资源”的政策要求，同时也顺应了大数据时代和人工智能兴起的时代背景，对教育教学工作者、教师、学生都产生了不可低估的作用。本文将从自动评分系统发展概述、自动评分系统信效度研究、自动评分系统应用三个方面展开论述。

关键词

自动评分系统，人工智能，英语写作

Review of Research on English Writing Automated Scoring System

Xinji Wang

School of Foreign Studies, Shanghai University, Shanghai

Received: December 24, 2025; accepted: January 16, 2026; published: January 28, 2026

Abstract

Due to the increasing number of large-scale exams and the growing number of test takers, traditional manual grading is no longer able to meet the demands of large-scale exams across regions and disciplines. Manual grading not only imposes a significant burden on grading teachers but also greatly reduces the efficiency of grading work. With the continuous maturation of natural language processing technology, automated scoring systems are expected to provide hope in addressing this challenge. Supported by cloud technology, automated scoring systems in modern educational tech-

文章引用: 汪鑫济. 英语写作自动评分系统研究概述[J]. 现代语言学, 2026, 14(2): 13-20.
DOI: 10.12677/ml.2026.142108

nology, based on corpora, score and provide feedback on students' answer sheets. This not only aligns with the policy requirements of "emphasizing the application of modern information technology and enriching English course learning resources" but also meets the era background of big data and the rise of artificial intelligence, which plays an undeniable role for educational and teaching staff, teachers, and students. This research discusses three aspects of automated scoring systems: an overview of their development, research on their reliability and validity, and their specific applications.

Keywords

Automatic Scoring System, Artificial Intelligence, English Writing

1. 引言

21 世纪初，随着信息技术的突飞猛进，大规模语言考试逐步由纸笔转向基于计算机和网络的考试。近些年在大数据和人工智能技术的辅助下，语言测试领域也开始尝试在命题、评分和反馈等环节中更全面地应用先进技术。目前，人工智能在教育领域的两大用途是学习和测评领域：在学习领域，逐渐开发了诸如自适应学习、机器翻译、自动反馈、自动情绪识别、等技术；在测评领域，逐渐开发了诸如自动生成考试、自动评分、智能监考等技术。在众多研究领域中，自动评分技术迎来了从 20 世纪 90 年代的光标阅读机读卡、21 世纪的主观题在线评分系统直到近十年出现的自动评分系统的重大革新。

以往大规模的考试评分仍然采用传统的人工模式，即根据事先拟定好的参考答案对不同的答卷进行评分，如果说客观题的阅卷较为简单，那么主观题的阅卷则需要依赖于评卷老师的主观印象，这在某种程度上使得主观题的评阅效率、评阅水平、评阅机制受到各类因素的不利影响，加之阅卷工作量带来的阅卷压力，评卷老师很难保证维持客观、公正的阅卷水准。因此，自动评分系统的研究应运而生。利用计算机进行英语作文评分和反馈的研究主要包括三种类型：第一种是以计算机和网络为平台的人工作文评分，王跃武(2004: pp. 67-73)致力于建立一种依托计算机及网络的高信度的大学英语四、六级考试作文网上阅卷管理系统。系统会向阅卷人随机分发试卷，并对阅卷行为、阅卷过程、阅卷质量进行监控[1]。在他的研究中，计算机技术只是作为作文评分的一个平台，阅卷工作仍交由人工处理。第二种是计算机辅助作文评估和反馈，曾用强(2002) [2]提出了基于语料库的过程化写作评估模式，并领导团队开发了过程化作文评估系统(PWESys1.00)。相对于第一种而言，它的系统更多地利用了计算机的词频统计功能，但依然停留在统计学层面。第三种是以梁茂成(2006)为代表的计算机作文自动评分，充分利用了计算机的词频统计和大规模运算能力，采用回归分析的方法，最后得出的自动评分结果与人工评分具有较高的相关度[3]。

纵观国内外文献，对自动评分系统的研究取得了实质性的进步，并已进入实用阶段。本文将从自动评分系统发展概述、自动评分系统信效度研究、自动评分系统应用研究三个方面展开论述。

2. 自动评分系统发展概述

对自动评分系统(automated writing evaluation, AWE; automated essay scoring, AES)的相关研究最早始于 1966 年的美国学者 Ellis Page (作文自动评分之父)，自此之后，相关研究领域和范围逐步拓宽并取得

了相应的发展，但大多集中在国外。早年间，针对英文短篇写作等作文题型的技术主要是以统计性为主，类似于评分类的短篇文本主观题(例如名词解释、简答及论述等)则通常会使用基于自然语言处理方式的相关技术手段。国外较早地采用了针对主观题应用的自动评分系统，并逐步推出了 PEG、E-Rater 以及 My Access!和 Criterion 等英文写作阅卷系统，纵观其历史发展，可以分为以下三个阶段：

第一阶段是上世纪 60 年代，由美国学者 Ellis Page 等人研制了首个自动评分系统 PEG (Project Essay Crader, PEG)，应用于大型考试的写作评分，其目的就是使大规模作文评分更加实际而高效。PEG 系统基于两个重要概念：目标变量(trin)和相关变量(proxes)。目标变量指的是人可以直接看出但计算机无法直接得到的变量；相关变量指的是计算可以直接得到的，并被用来估计目标变量。通过自然语言处理技术，PEG 从文章中抽取可以直接反映文章质量的表层语言特征，例如作文长度(流利度)、关系代词等数目(句子结构的复杂度)、句子长度变化(文本的措辞)。当这些表层语言特征被提取之后，再通过多元回归分析求得回归系数，以此来实现对文章的评价。但 PEG 提取文本的语言特征，只考虑文本的写作质量，而不考虑文本内容，无法对文本的内在意义进行评估，这也是 PEG 系统饱受诟病的地方之一。经过多年的发展，PEG 系统评价的有效性一直没有得到广大学者的认同。

第二阶段是上世纪 90 年代，由美国教育考试服务中心(ETS)开发的 E-rater 系统以及 Thomas Landauer 等人研发的 Intelligent Essay Assessor (IEA)系统。如果说 PEG 系统从语法质量和词汇准确性的角度来对作文进行评价，那么 E-rater 自动评分系统则使用了一种基于隐语义的分析模型(Latent Semantic Analysis)，它不仅考虑了作文的形式，同时也考虑了作文的内容，从而进行整体评价。E-rater 系统提取了目标作文的语法、词汇用法、机制和风格四方面的特征，通过各项特征值表现了在相应方面出现的错误数。此外，E-rater 还通过自然语言处理提取作文主题、主要观点、支持观点等多方面因素来进行分析。在提取特征之后，通过采用多元回归的方法来预测作文最终的分数。但这一类系统侧重于评价文本的抽象概念，而忽视了对文本的文体选择、形式正确等方面的重视。

第三阶段是 20 世纪初期，由 Vantage learning 公司研制的 IntelliMetric 自动作文评分系统。与上述系统相比，IntelliMetric 提取的目标特征变量要多得多，包括语义学特征、句法特征、主题切合度等 300 多个变量。IntelliMetric 使用了非线性和多维度的学习模型，主要从 5 个维度来进行评价：文章的整体性、文章的逻辑组织、句子的结构、文章的中心思想是否符合题目要求、文章是否符合标准书面英语的规则。随着人工智能的不断发展，除了上述提到的评分引擎之外，还出现了许多诸如 My Access!、Holt Online Essay Scoring、Criterion、Writing Roadmap 等智能评分系统，他们不仅能为学生提供作文的基础得分，还能从文章的内容、词汇、语法等维度提供个性化反馈，目前一直为国外中小学课堂所采用。

国内的有关自动评分系统的研究尚处于起步阶段，目前具有代表意义的主要有：梁茂成教授主导的大型英语测试论文自动化评分系统、《新视野大学英语》论文自动化评分系统、“Bingo”英语论文智能打分系统、IN 课堂智能作文批改系统、科大讯飞等。目前，一些高校已经开始采用其中一些系统，但是大多数的应用还局限于有固定格式的语文及英文作文的自动阅卷流程，还未正式运用到大型作文评分、教学反馈和教学实践中。

3. 自动评分系统信效度研究

在语言测试中，对于作文部分评分的效度很高，但信度常常不高，尤其很难保证评分的客观一致性，因此，国外早期对于自动评分系统的研究焦点主要集中在信度和效度问题上。Burstein & Chodorow (1999)最先将 E-rater 应用到 GRE 和 TOEFL 考试中，并且比较了其外语作文在人工评分和自动评分上的差异[4]。研究表明，尽管 E-rater 在满分六分的评分范围内差异的绝对值不大，但它与人工评分仍然存在着统计上的显著性差异。Burstein 等人首先将实验对象分为二语学习者和母语学习者，并在人工评分和 E-rater

评分的框架下对他们的作文进行比较研究。研究虽然得出人工评分的均值与机器评分的均值差别不大，但该差异具有统计显著性($F = 5.469, p < 0.05$)，这说明其中仍然存在一些因素从而影响机器评分的准确性。

以上提到的系统中，除了 E-rater，都是针对母语写作，并不适合用于外语写作评分。然而，E-rater 也恰恰忽视了一个非常重要的因素，那就是低水平的英语作文中高频率出现的词汇和句法方面的错误。这也导致了“传统的 NLP 语法分析器在 EFL 的教学应用上，尤其是作文自动评分上至今尚未取得广泛的成功”(Lonsdale & Strong-Krause, 2003) [5]。为了弥补这方面的不足，Lonsdale & Strong-Krause 基于 Link Grammar (LG)的句法分析器来分析评判英语学习者的作文。LG 以链接语法理论为基础，以链接图(graph)为核心，由连接词与词之间的类型化链接组成。这种结构允许生成 HPSG 和依赖性结构等不同形式，超越了传统的句法解析方法。运用到英语作文中，LG 分析器能够将其中的单词构成有句法意义的词对，比如：助动词 + 动词、介词 + 宾语、动词 + 宾语等。但是由于只分析句法以及句法分析器自身的不完善，机器评分的准确率较低。

Elliot, S. & C. Mikulas [6]也对英语写作自动评分系统的信度进行了实证研究，得出结论：英语写作自动评分系统是一种有效的语言辅助手段和语言辅助工具。他们首先将人工评分与自动评分得到的数据进行了对比，并且观察了二者之间的拟合度和关联性，以此来分析自动评分系统评分的信度，并且肯定了自动评分系统对人工阅卷带来的诸多好处。然而，也有学者对于过分依赖自动评分系统表达了担忧。如 Reilly 的研究中表明，自动评分系统与在线人工反馈在一堂写作网络公开课中的应用存在很大的差异，因此对该系统的可信度和有效性产生了怀疑。由此可见，自动评分系统要想更好地在语言教学课堂上使用，还需要进一步改进。

相对于国外而言，国内的自动化评分系统的研究起步较晚，在最近的 15~20 年才有所突破。北京航空航天大学外国语学院梁茂成教授是中国学生英语作文自动评分系统研究的先驱，他率领团队开发了“大规模考试英语作文自动评分器”等相关软件，并发表了《中国学生英语作文自动评分模型的构建》等相关著作，为发展国内自主的写作自动评分系统打下了坚实的基础。梁茂成(2006) [7]首先比较了三种典型的自动评分系统：PEG、IEA 和 E-rater 的优势和劣势，并且给出了相关建议。此外，他还根据前人对于外语写作质量的研究，集中关注了三种系统中学生作文的语言(流利度、准确度、复杂度)、语法、组织等特征因素，并结合实验结果进行评估。研究表明，人工评分和自动评分之间存在的差异是显而易见的。自动评分系统相对人工评分可以有效降低个人主观因素的影响，提供更加客观、准确的评价结果。但自动评分系统也可能存在某些误判，尤其是语言使用的细节方面。因此，在自主研发自动评价系统的过程中，应从多角度考量评分标准，以确保评价体系和评价系统的信效度。

基于此，越来越多的国内学者将目光投向了自动评分系统的研究中。李艳和葛诗利(2008)对大学英语作文自动评分中分级词表的效度展开了研究[8]。如何选取一个描述清晰、刻画准确的大学英语作文分级词表一直以来都是自动作文评分中的核心问题。研究探讨了建设大规模大学英语写作语料库的可行性，以及现有词表的缺陷，提出了应适当减少 1 级词汇的数量，调整 2 级词汇数量，增加 3 级词汇数量的改进建议，并表明了只有利用有针对性的、高效度的分级词汇表获取词汇分布特征，才能取得一个较好的评分效果。黄红兵(2015) [9]以“句酷批改网”为在线作文自动评分工具，首先肯定了句酷网在词汇、语法两个角度提供的详细评价和反馈，但也指出其对冠词和介词的检测不够敏锐，同时也缺乏对作文内容的分析。研究还结合了计算机辅助大学英语教学原理和形成性评价理论，提出了“三阶九评模型”。即作文经过“初稿”、“修订稿”和“定稿”三个阶段。在前期，教师在自动评分反馈和同学评价的基础上做出综合评价，学生以同学评价为参考、以教师评价为准绳修改，最后得出定稿。该模型能够有效减轻教师作文批改负担、提高学生作文反馈积极性。

李艳玲和田夏春(2018) [10]选取了 2016 年 11 月“国际人才英语考试”(English Test for International

Communication, ETIC)中 645 篇写作语料，利用一致性方法(Consistency Estimates)和一致率方法(Consensus Estimates)对语料进行数据处理，再通过人工评分与 iWrite 英语写作教学与评阅系统 2.0 版本评分进行多角度对比。结果表明，iWrite2.0 机器评分几乎可与人工评分相媲美。张国强和何芳(2022) [11]利用 L2 Lexical Complexity Analyzer、L2 Syntactic Complexity Analyzer 以及 Coh-Metrix 软件分析对比了 826 篇大学生英语六级作文在词汇使用、句法复杂、语篇连贯和言语失误四个维度上的语言特征，探讨了语言量化指标与分数之间的关系，以此构建各自的打分模型，并对模型的预测变量及效度进行了对比分析。研究表明，自动评分系统在处理不同主题类型作文的评阅过程中呈现出不同效果：对名言警句型作文的评分效度较低，而对现象观点类作文的评分效度较高。

总的来说，不同类型自动评分系统各有优劣，虽然总体上人工评分与自动评分差异较小，但在实践中并没有被广泛运用到考试中，这也在一定程度上解释了为什么当前自动阅卷评分系统仍然未能有效地实现市场化运作，因此还需要进一步改进和研究。

4. 自动评分系统应用研究

自动评分系统的引入是否对课堂内学生的写作教学产生影响，即学生的写作水平是否得到了提高，国外的诸多学者展开了相关实证研究，如美国加利福尼亚州的初中和高中已经将 My Access 和 Criterion 应用在课堂上，采用对比的量化研究方法，并发表了相应的研究成果。

早在 2004 年，Eliot & Mikulas 推出了自动评分系统 My Access 并开展了一项关于中小学学生写作水平的实证研究。实验对象是使用该系统的美国学生，从而研究他们在美国州统考中写作成绩是否有所提高。研究表明，My Access 系统的使用对学生的写作能力有了明显的提升。My Access 为学生提供一个写作环境，储备了 200 多个备选写作题目，针对不同主题的作文能够给出实时评分与诊断性反馈，激励学生不断修改作文以提高写作能力。但由于其中的两项研究没有包含对照组，因此很难判断系统对写作结果产生的影响。此外，研究还发现无论他们最初的写作能力如何，都可以在使用该系统后获得显著的提高。但需要注意的是，由于获得的被试人数和被试对象的局限性，缺乏实验的科学性和严谨性，因此并不能适用于所有的情况。例如以英语为非母语的学生为例，从 My Access 提供的作文评估反馈来看，一个是提交的作文语言非常地道，这远非中国大多数中小学学生能够达到的标准，另外就是所反馈的语言风格也并不适用于中国学生英语写作的教学。

Shermis *et al.* (2004) [12]采取随机抽样的方式来确定实验组和对照组，研究发现州统考后两组学生的成绩并没有显著性差异。针对前人的研究，Warschauer & Ware (2006) [13]发现了其中的问题，那就是组内随机抽样所造成的误差；同时，他们也指出任何教学改革，尤其是牵涉新技术的应用，只有当其完全融入到课程教学中，才有可能产生相应的效果。Rich *et al.* [14]分别在 2008 年和 2013 年依次研究了 Writing Road Map2.0 (WRM2.0)系统对学生写作能力的影响研究。结果表明，在使用了 WRM2.0 软件之后，学生的考试成绩，尤其是写作成绩，得到了明显的提高，这有赖于作文评分系统的积极反馈。Jiang Yaoyi (2015)以中国学生的英语写作为例，探讨了目前主流的自动评价系统背后的算法模型，指出了不同评价体系所代表的不同评价方法，表达了自动评价系统对学生语言学习的正向影响，并通过观察参与者的学习行为来为今后的自动评分系统的研究提供了参考依据。由此可见，在写作教学中，自动评分系统经过一系列的优化和革新，已获得了一定的成效，逐渐受到广大学者的认可，并被广泛运用到课堂教学中[15]。

与国外相比，国内对自动评分系统的研究处于刚起步的阶段，从最初的偏重于介绍性、探讨性(参见韩宁，2009；陈潇潇&葛诗利，2008)逐步向应用性(参见唐锦兰&吴一安，2011；张双祥，2014)研究方向发展。

韩宁(2009) [16]首先介绍了目前在美国大规模考试和英语教学中最为流行的几个作文自动评分系统

的基本原理并进行简单述评。正如 Page (2003) [17]强调的那样，计算机并不能像人一样去评判一篇作文，因为计算机只是“编程让它做什么”它就做什么，而并不能像人一样去“鉴赏”一篇文章。韩宁指出，上述自动评分系统最大的批评仍然是它们只能按照程序去检查文章是否完成了规定的任务，因此，在市面上也诞生了一些培训机构声称他们可以向学生提供类似“how to fool the E-rater”之类的课程，学生通过掌握“获得高分”的诀窍，从而达到自己的目标。要想使得自动评分系统能够在大规模考试中使用，最稳妥、最有效的方式仍然是采用一个人工评分员和一个机器评分员协同工作的模式。陈潇潇和葛诗利(2008) [18]也分析了当前在作文评分技术及应用方面具有代表性的 6 种系统的长处与不足，并且指出当前我国随着大规模考试人数的不断增加，如中考、高考、对外汉语教学的 HSK、大学英语四六级考试等，导致教师评卷工作量越来越重，而且评卷信度饱受争议。从某种程度上说，自动评分系统的引入能够有望突破写作批改量大、难度大的瓶颈，为教、学双方带来切实的帮助，尤其是如何利用自动评分系统切实提高学生的实际语言应用能力仍然是未来研究的焦点之一。

唐锦兰和吴一安(2011) [19]首先回顾和分析了迄今为止国内外对英语写作自动评价系统的相关应用研究，研究发现，现代教育信息技术的引入已经不仅是技术层面的革新，它还是一场牵扯到使用理念、认识、方法和行为等方面的系统性变革。人工智能如自动生成考题或学习材料、自适应学习、自动评分和自动反馈等技术在教学中的应用越来越广泛，从某种程度上来说，传统的教与学模式已经不再满足时代需求，面对这一新生事物，我们需要认识它，进而把握它，在教学中扬长避短。张双祥(2014) [20]以冰果网为例，运用问卷调查、访谈以及对学生作文样本的分析等方法探索在线写作自动评价系统在大学英语写作教学中的应用效果。研究选取了某大学非英语专业二年级共 122 名学生参与，要求这些学生每周完成一篇作文提交冰果系统评阅。从问卷和访谈的结果可以看出，冰果智能评价系统能够让学生通过修改作文来降低语言错误，从而提高他们的写作水平。

近二十年来，得益于自然语言处理技术的巨大进步，语言测试中的主观题机器自动评分终于成为可能。自动评分技术从最初阶段(依赖于人工设计的规则和模版进行评分)，到机器学习阶段(研究人员开始使用分类器或回归模型进行评分)，再到近些年深度学习阶段(能够自主学习和抽取特征的神经网络模型，能更好地处理复杂的语言结构和语义关系)。AI 自动评分系统在这些大型标准化考试中的应用是语言测试领域里程碑式的成果。

自 2022 年 11 月，以 ChatGPT 为代表的生成式 AI 横空出世以来，众多学者开始积极探索这一前沿技术与英语写作评价实践相融合。当前，生成式 AI 在英语写作评价领域的应用已经取得了突破性的进展，为师生提供了全新的写作辅助和评估工具。

学生可以利用 ChatGPT 进行自我写作评价，通过自动化修正语言瑕疵和询问语法、语言、结构等问题来提升自己的写作能力。其次，生成式 AI 还能根据学生的作文水平和需求，提供个性化的反馈和指导。这一技术的融入不仅弥补了人工评价和以往 AES 在错误纠正、语言优化、文本分析上的不足，还能通过设计个性化的提示词，为作文添加论据和细节，提高作文的连贯性和完整性，从而极大地提升英语写作评价的质量(Feng & Zhang 2024) [21]。更重要的是，这一技术已逐渐被融入英语写作教学的各个环节中，如 Chen 与 Lu (2024) [22]基于活动理论框架，提出了 ChatGPT 融入大学通用英语写作教学的四大阶段即写作前准备、写作、评阅及反思。Mao (2024) [23]则深入研究了生成式 AI 在写作评价中的实际效果，研究表明，ChatGPT 辅助英语写作有助于改善目前“笼而统之”的反馈模式，有效提升写作反馈的信度和效度。同时，该技术通过增强人机话语互动，促进了学生语用能力的实质性提升。

但深度学习算法和自然语言处理技术的进展也给 AI 自动评分系统提出新的挑战。如何确保 AI 评分的可信度？如何提高评分质量？如何有效利用人工评分的语料数据？这些都将成为自动评分领域持续的研究热点。同时，已有研究关于如何确保 GenAI 在自动评分上的有效性和自动反馈上的准确性，以及有

效规避可能存在的机械性错误和偏见，尚缺乏深入探索。

总的来说，国外对于自动评分系统的研发和应用起步较早，且已有大量的文献和技术支持，形成了一套成熟的产业链；而国内对自动评分系统的研究起步较慢，在最近的20年才慢慢从原理、理论向应用研究转型，由于人工智能的不断演变，未来将会出现越来越多的人工智能辅助投入到教育人工智能产业的实践，应用于教育测评和教学实践中。

5. 结语

在当前背景下，人工智能无疑掀起了一股时代浪潮，传统的教育领域内那些复杂、繁琐及批量化的工作都可以由自动化程序来辅助完成。借助于自动评分系统技术的加持，对教师来说，学校教师能够从大规模的考试评阅工作中脱身出来，以便于更好地将时间和精力投入到日常教学及科研工作的精进上来；对学生来说，自动评分系统的实时评分和自动反馈技术能够为学生实现自主学习提供硬件及软件上的便利，增强学生的学习能动性；对学校来说，人工智能技术的引进符合国家政策要求、时代需要，数字化课堂、智能化教学能够为教学教务管理者、教师及学生不断赋能，走在时代的前列。

近年来，信息技术和人工智能技术的发展势头和应用范畴远超人们的预期。对此，走在信息和统计测量等技术应用最前沿的语言测试学者桂诗春教授早有预见，他在“语言测试：新技术与新理论”中提到“在测试中应用新技术必须因地制宜，能用到高技术固然很好，条件未具备时，低技术也未尝不可。但我们不应该强调条件，就安于无技术的状态”[24] (桂诗春，1989：p. 2)。因此，技术应用必将成为语言测试领域乃至整个应用语言学领域研究的重要方向。但发展教育人工智能依然要以教育理论和实践为主，专注开发和不断地打磨能真正解决教育行业测、评、教、学、练各环节痛点的技术，才能在教育领域掀起一场真正的变革。

参考文献

[1] 王跃武. 大学英语四、六级考试作文网上阅卷实验研究[J]. 外语界, 2004(5): 74-79.

[2] 曾用强. 过程化的写作评估模式[J]. 福建外语, 2002(3): 26-31.

[3] 梁茂成. 学习者英语书面语料自动词性赋码的信度研究[J]. 外语教学与研究, 2006(4): 279-286+320.

[4] Burstein , J. and Chodorow, M. (n.d.) Automated Essay Scoring for Nonnative English Speakers. https://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf

[5] Lonsdale, D. and Strong-Krause, D. (2003) Automated Rating of ESL Essays. https://aclanthology.org/W03-0209/

[6] Elliot, S. and Mikulas, C. (2004) The Impact of My Access! Use on Student Writing Performance: A Technology Overview and Four Studies. *The Annual Meeting of American Educational Research Association*, San Diego, April 2004.

[7] 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.

[8] 李艳, 葛诗利. 大学英语作文自动评分中分级词表的效度研究[J]. 外语与外语教学, 2008(10): 48-52.

[9] 黄红兵. 在线大学英语写作形成性评价模型构建研究[J]. 现代教育技术, 2015, 25(1): 79-86.

[10] 李艳玲, 田夏春. iWrite 2.0 在线英语作文评分信度研究[J]. 现代教育技术, 2018, 28(2): 75-80.

[11] 张国强, 何芳. 英语作文自动评分系统的信度和效度研究——基于不同类型写作任务文本量化特征分析[J]. 外语测试与教学, 2022(1): 44-56.

[12] Shermis, M., Burstein, J. and Bliss, L. (2004) The Impact of Automated Essay Scoring on High Stakes Writing Assessments. *The Annual Meeting of the National Council on Measurement in Education*, San Diego, April 2004.

[13] Warschauer, M. and Ware, P. (2006) Automated Writing Evaluation: Defining the Classroom Research Agenda. *Language Teaching Research*, **10**, 157-180. https://doi.org/10.1191/1362168806lr190oa

[14] Rich, C., Harrington, H., Kim, J. and West, B. (2008) Automated Essay Scoring in State Formative and Summative Assessment. *The Annual Meeting of American Educational Research Association*, New York, March 2008.

[15] Jiang, Y. (2015) An Automated Essay-Evaluation Corpus of English as a Foreign Language Writing. *British Journal of Educational Technology*, **46**, 1109-1117. https://doi.org/10.1111/bjet.12292

[16] 韩宁. 几个英语作文自动评分系统的原理与评述[J]. 中国考试(研究版), 2009(3): 38-44.

[17] Page, E. (2003) Project Essay Grade: PEG. In: Shermis, M. and Burstein, J., Eds., *Automated Essay Scoring*: *A Cross-disciplinary Perspective*, Lawrence Erlbaum, 43-54.

[18] 陈潇潇, 葛诗利. 自动作文评分研究综述[J]. 解放军外国语学院学报, 2008(5): 78-83.

[19] 唐锦兰, 吴一安. 在线英语写作自动评价系统应用研究述评[J]. 外语教学与研究, 2011, 43(2): 273-282+321.

[20] 张双祥. 大学英语写作教学中在线写作自动评价系统应用研究[J]. 当代教育理论与实践, 2014, 6(11): 100-102.

[21] 冯庆华, 张开翼. 人工智能辅助外语教学与研究的能力探析——以 ChatGPT-4o 和文心大模型 4.0 为例[J]. 外语电化教学, 2024(3): 3-12+109.

[22] 陈茉, 吕明臣. ChatGPT 环境下的大学英语写作教学[J]. 当代外语研究, 2024(1): 161-168.

[23] 毛延生, 王一航, 邢艳茹. ChatGPT 辅助高中英语写作反馈的实证研究[J]. 教育测量与评价, 2024(1): 3-13.

[24] 桂诗春. 语言测试: 新技术与新理论[J]. 外语教学与研究, 1989(3): 2-10+80.