# 自动作文评阅系统评分效度验证及教学启示[①]

◎王　建　张藤耀

## 1　引　言

写作是英语学习的重要组成部分。但写作评估费时费力，尤其在中国 EFL（English as a Foreign Language）教学环境下，大学公共外语教师一学期所教学生数量可多达数百名。限于时间和精力，教师们不得不减少学生的写作练习，以减轻作文评阅的繁重工作，结果导致学生写作训练机会减少，获得及时和详细反馈的机会减少，学生的英语写作水平不能得到实质性的提高。近年来，为解决这一棘手的问题，教育技术开发商基于文本分类技术、自然语言处理、人工智能和潜在语义分析的，在自动作文评分（Automated Essay Scoring，　如 PEG™，Intelligent Essay Assessor，IntelliMetric，Electronic Essay Rater 等）引擎发展的基础上，研发出作文自动评价（Automated Writing Evaluation，简称 AWE）系统（Deane，2013）。国外著名的 AWE 系统如 Criterion，MY Access! 及 WriteToLearn 已被广泛运用于写作教学中，不但增加了学生练习的机会，而且还在内容、组织结构、词汇和语法等方面提供及时、详细的反馈和指导，从而将系统的功用从纯粹的评分引擎转变为计算机辅助语言教学工具（Ranalli，2018；Sarré et al.，2019）。

中国 AWE 系统的研发相对较晚，但是近十年来，商业性的系统如批改网、iWrite、冰果智能评阅系统等已广泛运用在中国大学英语写作教学中。如开发商所言，机器评阅在及时性、高效性和客观性等方面具有优势，学生可利用系统提供的“支架性（scaffolding）”反馈激活相关的英语知识，从而促进学生二语的发展，教师亦可在最近发展区（Zone of Proximal Development）理论指导下，引导学生进行同伴反馈。（张珊珊、徐锦芬，2019）但值得一提的是，在机器评分的有效性和真实性仍然扑朔迷离的情况下，一些教师完全依赖系统对学生的书面产品进行评分，并将机器分数直接纳入形成性评估中，这极有可能导致公平性问题。此外，为了获取高的分数，学生倾向于迎合机器的评价标准欺骗系统，然而这些标准可能与人工评阅者的标准大相径庭，或与写作构念（writing construct）毫无关联。（Powers et al. 2002）

① 本文系四川省民办教育协会立项课题“教育信息化背景下民办高校英语专业写作教学有效性研究（MBXH19YB016）”及全国高校外语教学科研项目“教育信息化背景下大学英语写作教学模式的构建及其有效性研究（2019JX0014B）”的阶段性研究成果。

尽管国内开发者高度评价系统的可靠性，声称自动写作评阅系统能够实现语言、内容、篇章结构及技术规范四个维度的智能评阅，但这些系统打出的分数是否真实有效，机器分数与人工分数是否高度一致，尚未广泛引起研究者及英语教师的注意。因此，本研究对国内某写作评阅系统的评分有效性进行初步研究，并探讨相关的教学启示。

## 2　相关研究

### 2.1　AWE 效度研究框架

效度是心理测量学中的一个广义术语。效度最早反映的是测试所要测量的构念是否被测到一定的程度（Kelly 1927），后来该术语指测量工具或手段的有效性，即能够准确测出所需要测量的事物的程度。目前国内有关 AWE 系统的研究更多关注的是自动系统对课堂教学辅助作用，如自动反馈对提高学生写作水平的作用、学生利用系统反馈的情况或学生对系统使用的认知，鲜有研究者从事 AWE 的效度研究，这是国内研究不足之处，因为在投入使用任何工具之前，使用者一般都希望知悉该工具的可靠性及有效性。相比，国外对 AWE 系统的效度研究较多，涉及的范围也较全面。影响较大的是 Kane 构建的自动评分系统效度论证（validity argument）框架，包括四个维度：评分（scoring）、泛化（generalization）、外推（extrapolation）和影响（implication）。（Kane，2013；Elliot and Williamson，2013）效度论证的具体方面较广，从人机评分的一致性、机器评分的稳定性、机器分数带来的影响到机器评分带来的后拨效应（wash-back effect）不等，详细的论证框架见表 1。

表 1　AWE 系统效度论证框架

| 效度论证纬度 | 主要研究问题 |
| --- | --- |
| 评分 | 1. AWE 系统与人工评阅者所衡量的文章特征是否相同？<br>2. 作文的系统分数与人工分数是否一致？ |
| 泛化 | 1. 系统提供的写作任务是否充分表征写作构念？<br>2. 学生在完成类似的写作任务时，系统给出的分数是否相似？ |
| 外推 | 作文的系统分数与其他写作任务（如多项选择题等）的分数之间存在何种关系？ |
| 影响 | 1. 作文的系统分数能否充分预测课程表现并用于入学分级（placement）？<br>2. 对于具有同等写作水平但不同背景的学生产出的文章，系统是否给出相似的分数？<br>3. 考生是否利用与写作构念无关的策略来获得更高的系统分数？<br>4. 对于相似类型的写作任务，系统给出的分数是否不受时间影响而保持一致？ |

表 1 中的效度论证框架较为全面地概括了国外自动作文评分系统效度研究领域的主要研究方向，总体来说，系统的评分效度受到研究人员更多的关注。

## 2.2 国内外 AWE 评分效度研究

有关 AWE 的评分效度研究始于 20 世纪末，至今依旧受到国外研究人员广泛关注。国外研究者对于该领域的研究大多集中讨论人机评分是否相匹配。例如，Deane（2013）报告称，AWE 系统注重文章的结构、语言结构等浅层特征，鲜有提供关于文章论证或修辞有效性的直接证据，这与人工评阅者差异很大。在现有文献中，研究者检验 AWE 评分有效性最直接的方式就是比较自动评分和人工评分是否一致，且普遍采用量化指标，如相邻吻合一致率（exact-plus-adjacent agreement rate）以及皮尔逊相关系数 r。不同于国内写作考试（如全国大学英语四、六级考试），国外写作考试（如雅思、托福考试）写作分数一般低于 10 分，人机评分相差 1 分则相差一个等级，因此相邻吻合一致率主要计算系统评分和人工评分的分数差小于等于 1 分的文章比例。皮尔逊相关系数用于统计人机评分的相关程度，系数越大说明两者的分数越趋向一致。由于分制的原因，国外研究报道的相邻吻合一致率和相关系数普遍较高，如有研究报道 IntelliMetric 的相邻吻合一致率高达 97%，相关系数为 0.83。（Rudner et al.，2006）

尽管国外研究大多报道 AWE 系统效度、信度均较高，但由于多数结果由开发者提供，鲜有独立的学者给出，因此结果的真实性不得而知。国内某系统开发者也验证了其开发的 AWE 系统的评分效度，比较了 1456 篇 15 分制作文的机器分和人工分的结果，发现 92.03% 的作文的分数差在 3 分以内，换言之，其相邻吻合一致性在 90% 以上。但这一结果亦是由开发者提供，真实情况如何，有待独立研究的进一步证实。

国内大型考试中作文模块的评阅工作仍由人工评阅者完成，因此大多数研究者对系统的评分效度关注不多，更多探讨系统反馈对提高学生写作水平的作用。国内文献中只有为数不多的独立研究人员进行了此领域的探索。万鹏杰（2005）对某 AWE 系统的研究结果显示人机间的相关系数为 0.324，远远低于开发者提供的系数。何旭良（2013）对句酷批改网的评分效度进行了研究，结果显示系统分数显著高于人工分数。另外值得一提的是，两个研究的样本均太小，前者为 85 篇文章，后者仅为 30 篇，研究结果的可靠性难以保证。此外，随着自然语言处理等技术日积月累地发展，AWE 系统的评分效度也有可能随之提高，万鹏杰及何旭良的研究可能会低估机器的能力。而且两项研究都没有揭示人机评分差异的分布情况及相邻吻合一致性，因而在研究广度和深度上存在不足。李艳玲、田夏春（2018）以“国际人才英语考试”的 645 篇实考作文为研究样本对 iWrite 2.0 的评分进行了研究，结果显示皮尔逊相关系数（五分打分公式人机分数 r=0.566）、克隆巴赫系数（Cronbach's Alpha=0.721）、完全吻合率（38.45%）、完全及相邻吻合率（97.98%）和卡帕系数（0.3518）都较高，据此得出结论 iWrite 2.0 评分较为理想。然而，白丽芳、王建（2018）对某作文评分系统的评分有效性进行了详细研究，除了收集人机相关系数、完全及相邻吻合一致性，还使用了最大分数差，指出系统无法可靠地评阅大学英语考试作文，容易误判人工高分作文。为解释人机评分差异成因，该研究还收集了研究语料在词汇、句法、篇章及错误等方面的量化特征并分别对人

工、机器分数建立回归模型，结果表明系统评分效度低可能是因其内部缺陷所致，机器评分主要依据浅层文本特征，不能像人工评阅那样分析深层文本特征，机器无法真正阅读、欣赏和判断文章，并且在分析深层句型结构或词汇搭配方面的能力不足。

国内 AWE 系统开发者在不同场合多次提到系统在英语作文评阅方面十分可靠，因此大部分高校都将此类系统融入写作教学中。但是，一个不容忽视的问题是：这些自动系统打出的分数与人工评阅者给出的分数是否真的高度一致？现有的研究结果并不一致。因此本研究将对国内某 AWE 系统的评分效度进行验证，以丰富该领域的研究，并探讨研究结果对大学英语写作教学的启示。

## 3　研究设计

### 3.1　研究问题

本文旨在回答：

（1）作文机器评分与人工评分是否一致；

（2）AWE 系统是否会误判特定类型的作文。

### 3.2　研究样本

本研究通过分层取样的方法，从“中国学习者英语语料库（Chinese Learner English Corpus）”中抽取 150 篇大学英语四级作文作为研究样本，所有作文均有人工原始分，分数从 6 分至 15 分不等。因语料库中 1 到 5 分作文量较少，本研究不予抽取，选取的各分数段的作文数量比例与整个语料库相当（表 2）。抽取四级作文为研究样本的另一个原因是，大学英语四级考试为高风险考试，人工评阅者在阅卷前须详细解读评分标准并接受打分训练，评分过程会受到监督，因此人工分数相对客观、权威。四级作文评分标准将考生作文划分为 5 个档次：2 分档、5 分档、8 分档、11 分档、14 分档，每档之间相差三分。

表 2　四级样本作文各分数段分布

| 分数 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14、15 | 总数 |
|---|---|---|---|---|---|---|---|---|---|---|
| 占比 | 12% | 17% | 20% | 20% | 14% | 7% | 6% | 3% | 1% | 100% |
| 数量 | 18 | 25 | 30 | 30 | 21 | 10 | 9 | 5 | 2 | 150 |

### 3.3　数据收集及分析

将从语料库中抽取的 150 篇文章按序排列（4001 至 4150），并提交系统进行自动

打分。目前该 AWE 系统的打分公式可人为选择，为充分保证人机评分的可比性，本研究选择四级打分公式（即满分为 15 分）作为该系统的评分依据。然后将机器打出的分数输入 excel 表格，利用 excel 及 SPSS18.0 计算三大指标：最大分数差指人机分数差绝对值的最大值；相邻吻合一致率指人机评分差绝对值小于等于 3 的文章数量与文章总量之比（四级作文每个档次相差三分）；皮尔逊相关系数检验人机评分的相关性。前者值越大，机器评分效度越低；后两者值越高，机器评分效度越高。有关人工、机器评分的描述性统计数据也由 SPSS18.0 算出，显著性设为 p<.05。

## 4　结果和讨论

### 4.1　作文人机评分的一致性程度

表 3 显示，该自动评阅系统给出的最高分和最低分均低于人工分。配对样本 T 检验表明机器分显著低于人工分（p=.000<.01）。150 篇样本作文的机器平均分为 8.049，显著低于人工平均分 8.77 分（p=.000<.01）。

表 3　人机平均分比较（n=150）

| | 最小值 | 最大值 | 平均分 | 标准差 | t 值 | 显著性（双侧） |
|---|---|---|---|---|---|---|
| 人工分 | 6 | 15 | 8.77 | 1.95 | 3.601 | .000 |
| 机器分 | 5.1 | 12 | 8.049 | 1.72 | | |

表 4 列出了人机分数差及相邻吻合一致率的分布。所有作文样本中，人机分数差为零的作文数量只有 3 篇，即完全一致率仅为 2%，人机分数完全匹配度（exact agreement）较低；117 篇文章的人机分数差小于等于 3 分，即人机评分相邻吻合一致率为 78%。国外研究指出自动评分系统与人工评阅间的相邻吻合一致性基本上要达到 75%～80% 这个水平（Burstein et al.,2004），按照这一标准，机器评分满足了这一要求。其余 33 篇的人机分数差均超过 3 分，这些文章极有可能被机器误判。国外大部分研究结果的相邻吻合一致率较高，甚至高达 90%，完全一致性在 48% 到 58% 甚至达到 80%（Ramineni and Williamson，2013），而本研究得出的结果偏低，究其原因，可能是由于样本量或者作文分数档存在一定差异，国外研究的样本量更大，而且作文通常为 1 分一档，而四级作文为 3 分一档。研究表明，分制的不同可能会导致这一差异，分制越低，相邻吻合一致性往往会越高。（Ramineni and Williamson，2013）比如以 3 分制进行评分时，若人工评阅者给出的分数为 2 分，机器给的 1 分、2 分和 3 分都与人工分相邻吻合，所以人机间理论上可以达到 100% 一致。

此外，本研究与国内多数 AWE 系统的人机一致性研究结果也存在很大的差异，这

有可能是由于样本的差异以及所研究的系统之间存在的不同导致的。比如目前各个系统具体的评分过程、评分标准并不透明，也未见详细说明，各个系统是否采用同样的评分方式不得而知，这些方面需要得到进一步的澄清。

表 4　人机分数差及相邻吻合一致率分布

| 分数差范围 | 0 | 0～0.5 | 0.5～1 | 1～1.5 | 1.5～2 | 2～2.5 | 2.5～3 | 3～3.5 | 3.5～4 | 4～4.5 | 4.5～7 | 相邻吻合一致性 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 作文总数 | 3 | 26 | 20 | 21 | 21 | 12 | 14 | 5 | 8 | 5 | 15 | 117 |
| 占比（%） | 2 | 17.33 | 13.33 | 14 | 14 | 8 | 9.33 | 3.33 | 5.33 | 3.33 | 10 | 78 |

人机分数差异较大的是编号为 4048、4110 和 4127 三篇人工高分作文，分数差分别为 6.4 分、7 分和 4.6 分。值得一提的是，四级作文的满分为 15 分，最大分数差如此之大，可见该自动评阅系统评分可信度需要引起使用者的注意。

本研究利用 SPSS18.0 对人机分数进行了相关分析，结果显示人机分数不显著相关，相关系数仅为 0.122（p=.136>.05，见表 5），而国外相关领域研究通常将相关系数设为 0.7（Ramineni and Williamson，2013），本研究结果远未达到这一起点值。研究结果的差异同样可能受分制的影响，也有研究表明不同分制的情况下，皮尔逊相关系数存在差异，分制越低，r 值越高。（Shermis，2014）然而，分制与系数的关系尚不明确，需要更多的研究证明。

表 5　人机分数相关性

| | 相关性 | | |
|---|---|---|---|
| | 样本总数 | r 值 | 显著性（双侧） |
| 四级作文 | 150 | .122 | .136 |

总的来说，该 AWE 系统的评分效度不尽如人意。描述性统计数据及三大效度指标都表明人机评分之间存在巨大差异，这就警示 AWE 系统开发者应着力提高机器的评分效度，同时教师应谨慎使用机器分数作为学业评估的一部分。

### 4.2　人机评分差异较大的作文类型及其成因

为进一步分析分数差的分布情况，本研究按照样本作文的人工分数将作文分为低（1～6 分）、中（7～9 分）、高（10～15 分）三类，统计分析显示三类作文平均分存在显著差异（p<.01），然后分别比较其相邻吻合一致性和平均分数差（见表 6）。结果表明，人机分数相邻吻合一致性在 6～8 分数段较高，为 92.85%；在 9～11 分数段为

中等，为 81.37%；12 ～ 15 分数段较低，仅为 36.95%。人机分数差的均值也随分数段的上升而上升，单因素方差（One-way ANOVA）分析显示，三类作文的分数差存在显著差异（p=.000）。事后多重比较分析（Post hoc Turkey's test）表明：低、中档作文的分数差不存在显著差异（p>.05），但均与高分档作文存在显著差异（p<.05）。不同等级作文平均分数差分布情况表明，该作文评阅系统有可能误判了人工判定的高分作文。

表 6　作文各分数段评分的一致性

| 分数段 | 低：6 ～ 8 分（n=73） | 中：9 ～ 11 分（n=61） | 高：12 ～ 15 分（n=16） |
|---|---|---|---|
| 相邻吻合一致性 | 92.85% | 81.37% | 37.5% |
| 人机分数差均值 | 1.43[a] | 1.88[a] | 4.11[b] |

注：人机分数差均值上标字母（如 a，b，c）相同表示不存在显著差异（p>.05），不同则存在显著差异（p<.05）。

国内外文献得出过类似的结论。如有研究比较了 E-rater（以 6 分制评分）和人工评分的一致性，发现在 5 分和 6 分两个高分档自动评分与人工评分的差异最大。（Burstein et al.，1998）也有研究指出 Criterion（以 6 分制评分）打出的低分比较可靠，打出的高分问题较大，并不能反映学生写作的真实水平。（Li et al.，2014）同样的现象在葛诗利、陈潇潇（2007）的研究中也有提及。

AWE 系统能够较为准确评价人工低分作文，可能主要是因为这类文章的语言和内容质量都较差。机器可以基于浅层的可量化的特征或语言错误给出客观的分数。在评价低质量的文章时，评分过程可以依靠可量化的特征或错误，但是在评分高质量的文章时，必须考虑文章的内容。由于机器无法理解一篇文章，它无法对文章的逻辑和思想做出任何判断，只能依靠一些可量化的特征来评估文章质量，而这些特征可能与一篇好文章毫无关联。（Condon，2013）因此这些量化特征可能不利于机器评分，进而导致对人工高分作文的误判。本研究语料来自大学英语四级考试作文，大学四级考试为高风险考试，在构建篇章时考生通常会选择简单的单词或常见的表达方式，少使用低频词或表达以避免出现错误，而这类文章有虽在词汇的复杂性方面较低，但文章的结构、逻辑、思想表达、语言的流畅度等方面可能做得很到位，人工评阅者在评阅作文时考虑的因素可能更为全面，不仅看词汇等浅层指标，还要考虑逻辑、内容等，但这些都是机器无法欣赏的，故有可能误判此类文章。限于文章篇幅，本研究并未对文章的量化特征进行统计分析，未来的研究可以利用语料分析软件收集样本作文在词汇、句法、篇章、错误等方面的量化特征，深入分析文本量化特征对人工评分及机器评分的影响及解释二者存在的差异。

### 4.3　研究结果对于大学英语写作教学的启示

不可否认的是，自动评阅系统可以为广大师生带来诸多便利。教师不用将大量时间用于评阅学生习作，而用于精心备课。AWE 系统打破了时空的限制，学生可以获得及时的写作反馈，学习自主性也可以提高。但是，本文通过定量分析的方法发现人机评分差异较大：机器分显著低于人工分，所有定量指标都不尽如人意；在无法理解文章内

容的情况下，机器极有可能误判人工高分作文。这与之前白丽芳、王建（2018）报道的系统评分效度存在的问题如出一辙。究其原因，目前用于机器评分的技术无法完全欣赏文章的逻辑、结构及修辞特征等方面。此外，目前机器仍无法与人工评阅者相比，人机评分所关注的方面可能存在差异，且评阅的方式也不尽相同，但存在何种差异需要更多的研究证明。因此，大学英语教师在使用机器分数时需要考虑到系统目前仍存在的种种缺陷。

必须指出的是，多数大学英语教师限于写作评估的压力将平时作文仅交由机器评阅，将机器分纳入学生最终的成绩中，但此时学生可能会质疑：机器分数是否真的可靠？仅利用自动反馈（缺乏教师反馈）是否真的有利于写作水平的提升？若学生作文仅由机器评阅，学生的写作热情无疑会受到影响。众所周知，目前机器还无法从真正意义上理解人的思维，还无法真正实现人机互动。最重要的是，目前自动系统评分的效度尚不明确，也并未引起广泛关注。如果机器分被纳入期末成绩中，有可能导致公平性的问题，因为机器可能会低估学生的写作能力，甚至误判高质量的作文。在英语写作教学中，各教师应合理运用机器评阅，可以采用人机结合的评阅方式，吸收二者评阅作文的优势。比如，限于技术的限制，目前机器评阅可以仅限于拼写、标点、大小写等技术规范方面，识别基本的语法错误（如主谓一致、冠词使用等）；教师应将写作视为真正意义上的互动交流，需要阅读学生写作的内容、结构、搭配、修辞等机器不太擅长的方面，给予学生写作建设性的反馈，给出适当合理的分数；教师还可以利用写作平台分配同侪协作的写作任务，相互给予深层次的交流与反馈，提高学生的写作积极性及写作兴趣。

总之，大学英语教师应顺应时代潮流突破传统的教学模式，但又不可完全依赖现代教育技术，不可完全忽视传统写作教学的作用，应在二者间寻求平衡。

## 5 结 语

本研究报道了国内某AWE系统的评分效度，结果表明该系统的评分效度相对较低，可能需要开发者进一步验证并不断提高。我们建议教育技术人员应与大学英语教师通力合作，进一步完善系统的评分机制，因为由于目前技术等方面的局限，教师还无法完全依赖机器。我们认为该领域需要引起国内更多独立研究者和使用者的注意。

本研究只是初步探讨了该系统的评分效度，还存在一些不可避免的缺陷。首先，较之国外同类研究，本研究样本数量相对较小；其次，未探讨人机在不同文本特征方面（词汇、句法、篇章等）对文章的评阅是否存在差异，因此对人机评分差异的解释深度不够；最后，未分析人机分数差大于三分的文章的特征，也未进一步研究可能被机器误判的人工高分作文在词汇、句法、篇章、错误等方面的特征。这些问题是未来AWE系统开发者和研究人员可以关注和解决的方向。但值得肯定的是，本研究对于写作教学融入AWE系统以及将机器分数纳入学生最终成绩起到了一定的警示作用，对大学英语写作教学有一定的参考价值。

## 参考文献：


[1] Burstein J, Braden-Harder L, Chodorow M, etal. Computer analysis of essay content for automated score prediction: A Prototype automated scoring system for GMAT analytical writing assessment essays [J]. *ETS Research Report Series*, 1998(1) : 1-67.

[2] Burstein J, Chodorow M, Leacock C. Automated essay evaluation: The Criterion online writing service [J]. *AI Magazine*, 2004, 25(3) : 27-36.

[3] Condon W. Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? [J]. *Assessing Writing*, 2013 , 18(1) : 100-108.

[4] Deane P. On the relation between automated essay scoring and modern views of the writing construct [J]. *Assessing Writing*, 2013, 18(1) : 7-24.

[5] Elliot N, Williamson D M. Assessing Writing special issue: Assessing writing with automated essay scoring [J]. *Assessing Writing*, 2013 , 18(1) : 1-6.

[6] Kane M T. Validating score interpretations and uses [J]. *Journal of Educational Measurement*, 2013, 50(1) : 1-73.

[7] Kelly T L. *Interpretation of Educational Measurements* [M]. Yonkers, New York: World Book Company, 1927.

[8] Li Z, Link S, Ma H, et al. The role of automated writing evaluation holistic scores in the ESL classroom [J]. *System* , 2014 , 44 : 66-78.

[9] Powers D E, Burstein J, Chodorow M,et al. Comparing the validity of automated and human scoring of essays [J]. *Journal of Educational Computing Research*, 2002 , 26(4) : 407-425.

[10] Ramineni, Williamson D M. Automated essay scoring: Psychometric guidelines and practices [J]. *Assessing Writing*, 2013 , 18(1) : 25-39.

[11] Ranalli J. Automated written corrective feedback: how well can students make use of it? [J]. *Computer Assisted Language Learning* , 2018(1):1-22.

[12] Rudner L, Garcia V, Welch C. An evaluation of IntelliMetric™ essay scoring system [J]. *The Journal of Technology, Learning and Assessment*, 2006, 4(4) : 3-21.

[13] Sarré C, Grosbois & M, Brudermann C. Fostering accuracy in L2 writing: impact of different types of corrective feedback in an experimental blended learning EFL course [J]. *Computer Assisted Language Learning*, 2019(7):1-23.

[14] Shermis M D. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration [J]. *Assessing Writing*, 2014 (20) : 213-246.

[15] 白丽芳，王建 . 人工和机器评分差异比较及成因分析 [J]. 外语测试与教学 , 2018(3) : 44-54.

[16] 葛诗利，陈潇潇 . 中国 EFL 学习者自动作文评分探索 [J]. 外语界 , 2007(5) : 43-50.

[17] 何旭良 . 句酷批改网英语作文评分的信度和效度研究 [J]. 现代教育技术 , 2013(5) : 64-67.

[18] 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(2) : 75-80.

[19] 万鹏杰 . 电子软件评估系统测试大学英语写作的研究报告 [J]. 外语电化教学 , 2005(3) : 11-13,31.

[20] 张珊珊，徐锦芬 . ZDP 视角下在线自动反馈对英语不同水平学习者写作的影响 [J]. 外语与外语教学 , 2019(5) : 30–39,148.

作者简介：

王建，男，1989 年 11 月生，四川成都人，英语语言文学硕士，西南交通大学希望学院讲师，主要从事英语写作的研究。

张藤耀，男，1986 年 12 月生，河南周口人，英语语言文学硕士，商丘学院，主要从事二语习得研究。