RESEARCH IN ENGLISH LANGUAGE EDUCATION JOURNAL

Volume 4, Issue 1 - August 2025
ISSN 3041-8909

Machine or Human? An Inspection of the Scoring of Writing Essays

Mostafa Ghaffari[1]

[1] *Corresponding Author*: Shahid Rajaee Campus, University of Farhangian, Qazvin, Iran.

Mostafa.ghaffari@gmail.com

ABSTRACT

Automated Writing Evaluation (AWE) systems are used to evaluate measurable characteristics of written texts, thereby creating a scoring model based on a compilation of essays. While considerable research has focused on the feedback provided by AWE systems, there is a conspicuous absence of studies examining these tools specifically in the Iranian context. Therefore, this research aimed to investigate the consistency of scores obtained from automated systems and human raters. Furthermore, it sought to explore the perceptions of EFL learners regarding the application of AWE in their writing practices. To facilitate this investigation, 30 male and female IELTS students participated, each writing two essays: one selected from topics provided by the AWE system and the other derived from Cambridge Official IELTS past papers. The essays were assessed by both My Access and three human raters. For the topics designated for the AWE system, a significant and robust positive correlation was identified between the ratings assigned by human raters and the machine. A similar significant and strong positive correlation was also found for the second essay, which did not utilize pre-defined topics. The results of two linear regression analyses demonstrated that the scores produced by the machine could significantly predict human scores for both pre-defined and non-pre-defined topics. Additionally, the findings indicated that My Access Home Edition is perceived to significantly enhance students' accuracy and autonomy, although it does not contribute to improved interaction. This study presents important implications for writing instructors and the field of second language education.

ARTICLE INFO:
Received: 2025-01-15
Reviewed: 2025-07-20
Accepted: 2025-07-20

Keywords:
automated writing evaluation, essay writing, human raters, machine raters, perception about AWE

1. Introduction

In contemporary discourse, there is a consensus that students require increased opportunities for writing practice, particularly in light of the advancements brought about by globalization (refer to National Commission on Writing in America's Schools and Colleges, 2003, p.3). Writing is undeniably a vital element in achieving proficiency

Citation: Ghaffari, M. (2025). Machine or human? An inspection of the scoring of writing essays. *Research in English Language Education Journal*, *4*(1), 118-136. DOI: 10.48210/relej.2025.18119.1118

in a language (Li, 2005). Furthermore, within the framework of English Proficiency assessment, writing serves as a significant evaluative measure (Rezaei & Lovorn, 2010). Nevertheless, the evaluation of students' writing and the provision of constructive feedback have consistently posed challenges for educators. The ability to write is recognized as a fundamental aspect of second language acquisition, which has been shown to be both difficult to master and complex to assess (Mehrani, 2017).

The history of Automated Writing Evaluation (AWE) programs goes back to the 1960s, when Page Essay Grade (PEG) was introduced. This program used a corpus of rated essays as a baseline to generate a scoring model. This was accomplished using multiple regression analysis on some characteristics of the texts that could be measured, like, the length of the sentences, average number of clauses, etc. (Shermis et al., 2001). The program stayed in focus until the 1990s when better possibilities were introduced by Artificial Intelligence (AI) for evaluation of writing (Warschauer & Ware, 2006).

Automated Writing Evaluation (AWE) programs possess the capability to substantially reduce the time and expenses associated with evaluating intricate skills such as writing (Weigle, 2010). Furthermore, engaging in iterative drafting and revision processes, coupled with formative feedback, leads to improvements in both essay quality and writing proficiency (e.g., McNamara & Allen, 2018). In 2016, a transition from score-centric feedback occurred, as third-generation AWE integrated guided activities generated through Natural Language Processing (NLP) techniques to enhance writing responses (Burstein et al., 2016). Knight and Buckingham Shum (2017) noted that this guided feedback technology is designed to foster individuals' growth and advancement over time, which is a fundamental aspect of formative automated assessment.

Numerous AWE programs are widely utilized, one of which is E-rater, currently employed as a secondary evaluator in the ETS TOEFL iBT independent writing task (Weigle, 2010). This indicates that E-rater may potentially supplant human raters in other high-stakes assessments as well. Consequently, its application must be validated against various criteria to gain acceptance and approval from test users and stakeholders. To this end, several validity frameworks have been established by prominent figures in the field (Chapelle, et al., 2015; Williamson, Xi, & Breyer, 2012), who have emphasized the necessity for further investigation into this matter. According to Weigle (2010), while automated scoring has simplified the evaluation process, there remains ongoing debate regarding the validity of these scoring systems, particularly in high-stakes or even low-stakes testing contexts. Concerns primarily arise from writing instructors, who argue that no machine can adequately assess the nuanced features that may be present in any written work. Clearly, empirical research is essential to address these concerns to provide a clearer understanding of the contributions of these systems.

Furthermore, an examination of the existing literature on the AWE phenomenon in Iran reveals a significant lack of research focused on the validation of such programs. Consequently, it is imperative to assess the validity of AWE, or Automated Essay Scoring (AES) programs, within a recognized framework in Iran. Research efforts in the country have generally taken a broad perspective on the impact of technology in writing, particularly regarding its contribution to writing quality. For instance, Tafazoli (2014) conducted a study demonstrating the effectiveness of Computer-mediated Corrective Feedback in English for Specific Purposes (ESP) courses, specifically in reducing grammatical errors through email in an English as a Foreign Language (EFL) context. It is also important to highlight that previous

investigations into AWE have predominantly employed quantitative methodologies; thus, there is a pressing need for more qualitative studies to provide a well-rounded understanding of AWE's validity in second language education. In light of this, the current study incorporated interviews as a qualitative component to adopt a mixed-methods approach.

To validate AWE programs, this study specifically explored the validity argument concerning the AWE program known as My Access. The researcher applied the framework established by Williamson et al. (2012) for the validation of these AWE programs. Additionally, the study sought to understand the perceptions of Iranian EFL learners regarding the role of AWE programs in enhancing their writing accuracy, promoting learner autonomy, and facilitating interaction. To gather data, the researcher utilized distinct questionnaires and conducted interviews as part of the qualitative research component. Accordingly, the following research questions were posed.

Research Question 1: Do scores obtained from MY Access and human scoring of defined and undefined essays significantly correlate?

Research Question 2: Are machine scores significant predictors of human rating scores? Do the prediction powers, if any, differ for defined and undefined topics?

Research Question 3: What are the perceptions of EFL learners as users of AWE about its effectiveness?

2. Review of Literature

2-1. Technology and Language Pedagogy

English language pedagogy in the classroom is fundamentally composed of three essential components: the educator, the student, and the English language itself. This pedagogical approach focuses on how educators can effectively support students in their acquisition of English, encompassing teaching strategies, instructional materials, and the activities that English teachers implement within the classroom. According to Brown (2014), educators should consider, "Your understanding of how the learner learns the language will determine your philosophy of education, your teaching style, your approach, your methods, and classroom techniques" (p. 7).

According to Colombi and Shleppgrell (2002), in the intricacies of the modern world, literacy encompasses much more than merely acquiring the skills to read and write for specific, isolated tasks. The ongoing evolution of technology and societal dynamics indicates that the nature of literacy tasks is also transforming. It is a well-established fact that numerous ESL and EFL learners utilize computers in their studies. Computer-Assisted Language Learning (CALL) pertains to the use of computers in the teaching and learning of a second language. Professionals in CALL create software and online resources designed to enhance second-language acquisition. Similarly, computer-assisted language testing encompasses various elements of language assessment and technology application, aligning with the framework for describing computer-assisted language tests as tools developed within CALT (Suvorov & Hegelheimer, 2014). Consequently, educators and language instructors regard computers as a vital component of pedagogy, serving as a facilitator for learning in the classroom (Chapelle & Jamieson, 2008, p. 2).

The application of technologies can facilitate the teaching of four language skills, including writing. Some authors (e.g., Graham,2020) believe that the best way to enhance writing skills is the application of a recurring model known as modeling-

practicing-reflecting cycles that focuses on the realization of instruction, practice, and formative feedback characterized by core features of writing. Along the same lines, Graham, Bañales, et al. (2020) assert that the best and most effective writing strategy initiatives are concerned with inducing in writers some plausible and purposeful procedures and "tools" which they can use in a wide range of writing tasks.

2-2. Automated Writing Evaluation (AWE)

A close look at the history of writing shows that writing technologies have transformed how individuals proceed with writing, as well as the instruction of writing (Graham, 2021). Indeed, new technological developments have contributed to the availability of language checkers, making the evaluation and provision of feedback on writing more rapid and personalized. One of these systems is automated essay scoring (AES). At the outset, automated essay scoring systems appeared as an evaluation tool through which writing was assessed through the implementation of large-scale standardized tests; however, a large number of these systems have undergone some modifications to fit classroom application. For example, they have been expanded to provide formative feedback. Regarding the automated writing evaluation (AWE) systems, learners can enter several writing cycles, including being exposed to feedback and writing revision. Indeed, in automated writing evaluation (AWE) systems, students are provided with ample opportunities to receive feedback. The provision of feedback on the learners' writing has been examined, with many scholars focusing on various aspects, forms, and strategies involved in providing such a response (Bitchener & Storch, 2016; Boggs, 2019).

According to McNamara (2022), consistent with interdisciplinary perspectives on writing aimed at supporting learners' writing in the classroom, AWE needs to uphold a community of learners, interlacing reading and writing instructional activities along with feedback to use reading and writing strategies. To this end, AWE systems have become more sophisticated since the 1960s. Today, Modern automated writing evaluation (AWE) systems are more sophisticated, so they often consist of versatile tools, including spelling and grammar checking ones. They are also equipped with services such as Grammarly, which can help the evaluators greatly, helping everyone to be a great writer (Koltovskaia, 2020). These sophistications, along with the emphasis placed by educators on the use of technological advancements during the last two decades, have now permitted AWE programs to gain commercial validity.

2-3. MY Access!

MY Access! (developed by Vantage Learning) is one of the widely-used scoring programs that is popular for its strong database. Using an AI-powered scoring machine, called IntelliMetric!, it scores any essay by comparing about 300 semantic, syntactic, and discourse-specific characteristics of it with the samples, archived on its database bank, which were previously rated by humans (Elliot, 2003). The output includes a holistic score of the essay on a 1-6 or 1-4 scale and generic feedback generated based on the grade level, detected genre, and the estimated score. Generic feedback is also provided based on grade level, genre, and score. It is also featured by a tool, called *My Editor*, that provides detailed feedback on text features like spelling, grammar, and the proper use of vocabulary.

MY Access! has come to be known as an educational composition tool that allows students to enhance their writing knowledge and skills embedded in an e-portfolio-

based setting. Within this context, instructors can write an essay assignment out of many unique prompts that cover a wide range of topics and text types, such as expository and narrative genres. Aimed at generating an integrated composition tool, the prompts are concerned with the main textbook series and already set standards. Indeed, these prompts provide the learners with rich chances to write in a cross-curricular manner with a focus on the subject areas, including experimental science, mathematics, and social science. Besides the subjects provided in MY Access!®, teachers are all able to contribute their essay themes to the system.

Instructors can lead the learners through pre-writing tasks and allow them to review model papers that are consistent with the prompts provided in MY Access!®. Learners can receive a wide range of feedback from the system while writing. They can also receive feedback based on the score they are given about the MY Access!® rubric. Having submitted an essay, the learner is provided with prompt feedback from IntelliMetric®. He/she can also receive from the instructor.

MY Access!® serves as an evaluative tool that gives both a holistic score and an analytical one in the domains of semantics, the development of content, heretics, linguistic devices, language style, and use, and Mechanics and Conventions. For the learners whose native language is Spanish or Chinese, feedback can be given in their native language if a teacher finds it necessary. Each learner is allocated an online portfolio using MY Access!®, so that all the composing drafts, grades assigned, editions, teachers' comments, reflective journal entries, and IntelliMetric® feedback can be accessed at any time. Moreover, instructors and administrators can access these portfolios at different levels and from any place, including class or home. Besides the comments, grades, and entries provided in the online portfolio, MY Access!® gives additional writing guidelines and tools.

Learners can proceed with learning how to write through practical writing exercises and projects. Empirical research reveals that the extent to which students proceed with writing has a positive correlation with writing ability (Chircop 2005). Indeed, studies show that intensive writing courses, which entail the composition of several drafts, as well as a high level of writing practices, including creating writing portfolios or projects to enhance successful writing, contribute greatly to promoting the effectiveness of writing aptitude among learners (Chircop, 2005). As pointed out by Reeves (2007), as long as learners invest in writing frequently, their capability of thinking, reasoning, analyzing, communicating, and performing on tests would improve to a great extent. Indeed, writing plays a pivotal role in the realization of learners' achievement. Effective educational programs perform periodical assessments, providing learners with many opportunities to be successful. It is worth noting that highly effective educational programs are characterized by a formative assessment program aimed at assessing writing performance (Reeves, 2007). MY Access!® gives the learners ample opportunities to compose and receive feedback much more regularly compared to the classic writing methods. Combined with a well-structured curriculum, the formative assessment provided by MY Access!® contributes to effective achievement.

Studies also reveal that effective and timely feedback plays an important role in enhancing learners' writing skills. The research results show that the regular provision of feedback in the early stages of writing courses induces in students a positive attitude toward feedback, which positively influences the quality of the writing (Cowie 1995). When learners earn a score of 2 on their essays, their motivation increases to a great

extent to submit a refined version of the essay. This is because such an improved version enables the students to progress to the next stage. Such a type of feedback enables the learners to figure out the constituent components of quality writing. Indeed, as pointed out by Reeves (2007), the lack of such immediate feedback turns testing into an ineffective "academic autopsy" which provides no opportunity for remediation. MY Access!® paves the way for the invaluable and appropriate feedback required for enhancing student writing accuracy.

In his meta-analysis, Marzano (2001) provides a comprehensive set of core research-based qualities for effective teaching of writing. Besides timely feedback, another important element is the use of clearly stated learning goals. MY Access!® is composed of thorough scoring rubrics accompanied by comments on recently composed papers in such a way that learners know what they should do to realize each one of the learning objectives. Teachers would do well to select plausible instructional goals, such as: "You are required to submit at least three drafts to the prompt and obtain a mark of at least 5 out of 6 on the ultimate submission." To give immediate feedback to learners, MY Access!® makes use of IntelliMetric®, Vantage Learning's automated essay scoring system. Learners can modify their composition by receiving feedback and resubmitting for the analysis of the essay. Research shows that these writing phases, namely, the reception of feedback, continuous revision, and the reception of more feedback, have proven to be vital for the enhancement of writing proficiency.

2-4. Intelligent Essay Assessor

Beginning in 1995, the Intelligent Essay Assessor (IEA) was created with the intention of evaluating—and perhaps even fully automating the grading of— written constituents across various disciplines, including social science, psychology, and language education (Foltz, Laham, & Landauer, 1999) of middle school- and undergraduate-level students. The eventual intent was to construct automated replies to evaluative writing exercises. In the past, IEA was utilized for middle grade and collegiate composition evaluation; however, it is now utilized for formative writing assessment testing. Additionally, it is utilized for the evaluation of new GED and science assessments. If implemented within the framework of high schools, the IEA may also be utilized for writing assessments to determine placements.

Assessment of learners' comprehension and ability to compose essays is made simpler by the IEA. This is due to the identification of the factors associated with the writing elements and constructs. These include: vocabulary resources; grammatical structures, rhetoric, cohesive devices, organization, and content. According to Landauer et al. (2011), vocabulary resources are evaluated in terms of the level of development that has been achieved in terms of age and also in terms of their lexical diversity. In addition, other analytical systems are moving toward investigations that employ advanced computational techniques to appropriately locate and identify e-text indicators that aid in assessing grammatical structures and other mechanical elements.

When it comes to the analysis of mechanics, other elements are employed to assess the correct spelling and punctuation rules. Content-oriented factors are assessed through latent semantic analysis (LSA), which is a statistical modeling tool that resorts to a voluminous pool of lexical items to model words used in a specific domain (Landauer et al., 2001). The new technologies help to analyze the content aimed at assessing aspects, including concepts, coherence, and the effectiveness of text summaries compared to gold-standard texts (Foltz et al., 2000). IEA adopts a machine

learning-oriented approach to determine the effective set of elements and the weights assigned to each one of the elements aimed at optimally modeling the grades for each essay. Based on these comparisons, it is possible to develop scoring models through which the scores can be predicted given new responses. Given the type of writing activity, it is possible to expand these models for rating writing performance. Following this scoring model, it is possible to promptly score newly written text through the analysis of the traits weighted. Moreover, the IEA has some specific guidelines for teachers, identifying off-topic essays.

3. Theoretical Framework

Over the last four decades, numerous studies have been undertaken to assess the validity of Automated Essay Scoring (AES) systems. For instance, Yang et al. (2002) identified three primary approaches in their research: First, a significant number of studies have focused on exploring the correlation between automated scores and human evaluations. Second, another group of studies has investigated the potential connections between AES scores and external benchmarks, such as instructors' assessments of students' writing abilities or results from other tests (Coniam, 2009; Vantage Learning, 2007; Weigle, 2002). The external factors analyzed include multiple-choice assessments, grading based on writing performance, instructor evaluations of student writing, and self-assessments by learners (Ben-Simon & Bennett, 2007). Third, the final approach has aimed to investigate the scoring mechanisms and cognitive frameworks utilized by AES systems. A notable concern in AES scoring pertains to the emphasis placed on essay length as a scoring criterion, as length is recognized as a significant predictor in the e-rater system (Chodorow & Burstein, 2004). According to Williamson et al. (2012), "agreement of automated scores with human scores has been a long-standing measure of the quality of automated scoring" (p. 8).

In the realm of validation, scholars such as Kane (2006) and Williamson, Xi, and Breyer (2012) have proposed a distinct conceptual framework for Automated Writing Evaluation (AWE). This framework encompasses several key components: explanation (including theme, activity, and scoring analysis), evaluation (the interplay between human and automated scoring), generalization (the degree to which findings can be applied to various tasks and test formats), extrapolation, and utilization (the application of scores and their outcomes). The Explanation component elucidates how four distinct AWE systems provide insights that facilitate the understanding of inter-system relationships and relevance construction. The Evaluation section offers comprehensive descriptions of the scoring mechanisms and evaluative functions inherent in AWE systems. The Generalization segment draws upon empirical studies utilizing AWE to assess learner performance. The Extrapolation component justifies the development of AWE systems that evaluate writing tasks beyond traditional academic essays. Lastly, the Utilization section addresses the role of AWE in decision-making processes. Xi (2010) has also raised a series of validity-related questions regarding automated scoring, which include concerns about the accurate representation of constructs and writing activities, the validity and reliability of scoring, and whether a test taker's awareness of scoring algorithms might affect their writing performance.

In their empirical investigation into the challenges encountered by children during reading, Graham et al. (2020) discovered that those struggling with reading difficulties achieved lower scores on essay assignments. Conversely, norm-referenced assessments, such as standardized tests, appeared to allow for a greater number of

mechanical errors compared to the writing evaluations created by researchers. Additionally, a study by MacArthur, Philippakos, and Graham (2016) indicated that participants who prioritized conventions and mechanics tended to be less proficient writers than those who recognized the importance of content and structure. In a meta-analysis of research on Automated Writing Evaluators (AWEs), Strobl et al. (2019) identified 90 different evaluators that employ various algorithms, including Natural Language Processing Systems (NLPS), to provide both summative numerical scores and qualitative feedback on essays and other open-ended responses. Furthermore, Daghbandan (2015) assessed the effectiveness of Grammarly Software in enhancing the writing accuracy of English as a Foreign Language (EFL) students. Similarly, Ghasemzadeh and Soleimani (2016) explored the impact of feedback from both Grammarly Software and teachers on the acquisition of passive verbs among Iranian EFL learners. However, it can be stated that there is a scarcity of empirical studies investigating the effects of AWE programs, such as My Access!, on Iranian EFL learners. To date, no research has examined the correlation between the defined essay scores from MyAccess and those assigned by human evaluators. Additionally, the perceptions of EFL learners in Iran regarding AWE systems remain unexplored. Moreover, on a global scale, most relevant studies have predominantly employed quantitative methodologies. This study aimed to address this gap by delving into the research questions mentioned above.

4. Methods

4-1. Research Design

This research took a quantitative design to answer the research questions. The first two research questions were probed through correlational analysis, and the third one was investigated by looking into the participants' answers to the questionnaire.

4-2. Participants

This study sought to assess the distinctions between human and machine essay evaluators, as well as to explore students' self-reported views on the implementation of Automated Writing Evaluation (AWE) systems. The research included thirty Iranian candidates preparing for the IELTS, consisting of fifteen males and fifteen females, all aged between 18 and 25, who were participating in an IELTS preparation course led by the researcher. All participants were at least at an intermediate proficiency level. To ensure uniformity among participants, a placement test was conducted, employing the English Ultimate course book published by Cambridge University Press.

Moreover, sixty essays authored by students were coded and printed for evaluation, subsequently being assessed by three individuals. These raters, who were non-native English as a Foreign Language (EFL) instructors, possessed Master's degrees in Teaching English as a Foreign Language (TEFL) and had approximately eight years of experience teaching various examinations, including IELTS, TOEFL, KET, PET, FCE, CAE, CPE, GRE, and others, both in Iran and internationally. The raters were Master's and PhD graduates in Applied Linguistics (TEFL). Prior to initiating this segment of the project, the researcher needed to confirm that the three raters fulfilled specific criteria. To this end, Williamson's (2013) framework for *Argument for Essay Scoring with Human and Automated Scores* was employed. This framework stipulates that raters must possess appropriate qualifications and complete a

training course. Consequently, three raters with Master's degrees in Applied Linguistics, who had adequate experience in essay evaluation and had instructed essay writing to diverse EFL learners, were selected for the scoring session. Before commencing the actual scoring process, the raters collaborated to establish the guidelines for the holistic scoring rubric utilized in MY Access. Additionally, they examined several essays that had been previously scored by MY Access to gain insight into how IntelliMetric assessed these essays in relation to the rubric's criteria.

4-3. Data Collection Instruments

Various research instruments were employed in the present study based on the validity arguments for AWE in the literature. The instruments were as follows: MY Access Home Edition, and students' self-reports questionnaire regarding their perceptions of AWE.

In this study, the researcher used MY Access! Home Edition, powered by Vantage Learning, to carry out this study. The cost of one year's access to this tool was 100 dollars. The Home Edition should not be considered an independent writing curriculum. Indeed, this system does not fashion the writing process in a classic, linear way. Also, it does not deal merely with the mechanics of writing, as is the case with other electronic tools. This system was not designed to substitute instructors altogether! But it provides users with a real, customizable learning setting.

At the heart of MY Access! is IntelliMetric®, which is an artificial intelligence scoring tool. It immediately assesses learners' writing based on a standard rubric, providing the learners with suggestions to enhance the quality of their writing. The feedback is consistent with the main traits of writing:

- Language Use,
- Focus,
- Organization,
- Grammar,
- Content Development,

Regarding this intelligent component, MY Access! includes instructional segments for each genre consistent with the age range, a detailed composition manual, organizational resources, and a set of word-processing means. Families and learners decide which features of the program fit them at each stage of the learning process.

In addressing the third question concerning students' perceptions, the study examined the views of Iranian EFL learners regarding the impact of AWE programs on their writing accuracy, learner autonomy, and interaction. This assessment was conducted after the learners had their essays evaluated and received feedback through MyAccess on five separate occasions. To gather data, the researcher employed distinct questionnaires. These instruments, initially created and utilized by Wang, Shang, and Briody (2013) in a Taiwanese study, had already undergone validation for use in an EFL context. Each section focused on accuracy, autonomy, and interaction and contained ten questions, resulting in a total of thirty items. Due to the limited number of participants, the construct validity analysis was not applicable for the context of the study. However, three experts in the field evaluated the questionnaire by filling out the content validity index (CVI) form, and no unnecessary item was spotted by them. Moreover, the reliability of the collected answers was estimated through Cronbach's alpha formula, and an overall acceptable reliability index of .73 was obtained.

4-4. Data Collection Procedure

Sixty student essays were evaluated by three human raters alongside a web-based machine rater known as MY Access! Home Edition. The essays focused on a persuasive topic: the comparison between Internet classrooms and traditional classrooms, chosen from a selection of 90 topics available within the software. The essays were composed in Word documents and subsequently emailed to the researcher. Each participant wrote 30 essays on a predetermined topic for the machine and 30 essays on a topic that was not defined. The latter topic was selected from the Cambridge IELTS 12 general training practice test. Both the machine and the human raters assessed the essays, and the resulting scores were correlated to address the first research question.

The assessment framework utilized by MY Access is based on a comprehensive rubric available on its website. This rubric features a scoring range from 1 to 6, defined by several essential writing characteristics. Each score is paired with a concise description of the relevant traits. For instance, to attain a score of "6," an essay must articulate and maintain "an insightful controlling or central idea and demonstrate a thorough understanding of the purpose and audience [in terms of Focus and Meaning]" while also "exhibiting minimal errors in paragraphing, grammar and usage, punctuation, spelling, and mechanics [in terms of Mechanics and Conventions]." In contrast, a score of "1" indicates an essay that fails to effectively communicate the writer's message and lacks adequate support for its ideas through details and/or examples (in Content and Development). In addition to the scores, the study also investigates the feedback provided by the AWE system and human raters. The analysis includes the following aspects:

Specificity and Detail: The feedback from the AWE system and human raters is compared in terms of specificity and detail. This involves examining how well each source identifies and comments on various aspects of writing, such as grammatical errors, sentence structure, argument development, and overall coherence.

Constructiveness: The study assesses the constructiveness of the feedback by evaluating the extent to which it helps students improve their writing skills. Human raters' feedback often includes actionable suggestions, while the AWE system's feedback is analyzed for its ability to provide meaningful and personalized guidance.

Clarity and Understandability: The clarity and understandability of the feedback from both sources are compared. The study examines whether the feedback is communicated in a clear and comprehensible manner, considering the students' level of understanding and familiarity with the terms used.

Perceived Usefulness: Participants' perceptions of the usefulness of the feedback are gathered through a follow-up questionnaire or interviews. This step helps to understand how students view the feedback's impact on their writing accuracy, autonomy, and overall interaction with the system.

The second objective of this study was to explore the ability of machine-generated scores to predict human rating scores. Two linear regression analyses were performed using machine scores derived from evaluations made by human raters. The predictive capabilities were subsequently compared using a Z test. To address the third research question, the study examined participants' perceptions of machine scoring. Participants were instructed to compose essays in response to five selected essay prompts from MY Access over five consecutive weeks. Their essays were scored and returned via email, along with feedback from MY Access.

After this process, participants completed a questionnaire designed to assess their perceptions of MY Access Home Edition. This questionnaire aimed to gather insights regarding participants' views on their writing accuracy, autonomy, and interaction.

4-5. Data Analysis

In order to answer the first research question, a correlational analysis was carried out. The second research question was probed through running two separate linear regression models and comparing the results through a z-test. Finally, in order to answer the third research question, participants' answers to the questionnaire were tested through a one-sample t-test.

5. Results

To answer the first question, correlational analyses were needed. The descriptive statistics of the scores obtained from the three raters and the machine are presented in Table 1.

Table 1

Descriptive Statistics of the Scores Obtained from Raters and Machine (N = 30)

		Minimum	Maximum	Mean	SD	Skewness Ratio
Defined Topic	Rater1	2.90	5.70	4.5167	.72019	-0.87
	Rater2	3.10	5.90	4.6200	.66405	-0.88
	Rater3	2.70	5.90	4.5533	.75873	-1.46
	Raters'	2.90	5.70	4.5633	.69389	
	Mean					-1.31
	MY Access	2.80	5.70	4.4933	.75381	-0.97
Undefined Topic	Rater1	3.20	5.80	4.4700	.77109	-0.19
	Rater2	3.40	5.60	4.4933	.74414	0.28
	Rater3	3.20	5.90	4.3733	.82752	0.74
	Raters'	3.30	5.73	4.4456	.75777	
	Mean					0.26
	MY Access	3.20	5.90	4.4033	.78892	1.31

As reported in Table 1, the given scores by raters and the machine had close means for both defined and undefined topics. Moreover, the skewness ratios were all indicative of normal distributions as they all fell within the legitimate range of ±1.96. Considering the normalcy of distributions, to answer the first research question, two sets of parametric Pearson correlations between raters' mean scores and the scores given by the machine were run. The results are presented in Table 2.

Table 2

Pearson's Correlation between The Mean of Human Raters' Scores and Machine Scores: Defined and Undefined Topics

		Raters' Mean	MyAccess
Raters' Mean (Defined topics)	Pearson Correlation	1	.972**
	Sig. (2-tailed)		.000

Raters' Mean (Undefined topics)	Pearson Correlation	1	.474**
	Sig. (2-tailed)		.008

** Correlation is significant at the 0.01 level (2-tailed)

The analytical findings presented in Table 2 indicate a significant and positive correlation between human ratings and machine ratings for defined topics, with a correlation coefficient of r = .972, n = 30, p < .01. This reflects a very large effect size (r^2 = .945). In contrast, for undefined topics, human ratings also show a significant and positive correlation with machine scores, yielding a correlation coefficient of r = .474, n = 30, p < .01, which corresponds to a medium effect size (r^2 = .225). To explore the second question, two linear regression analyses were conducted.

The results illustrated in Table 3 reveal that machine scores account for 94.5 percent of the variance in EFL human rating scores for defined topics (R = .972, R^2 = .945), while they explain 22.5 percent of the variance in EFL human rating scores for undefined topics (R = .474, R^2 = .225).

Table 3

Model Summary (Defined and Undefined Topics)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.972a	.945	.943	.16620	2.392
2	.474a	.225	.197	.67894	1.963

Model 1: Defined Topics; Predictors: (Constant), Mean; b. Dependent Variable: MyAccess
Model 2: Undefined Topics; Predictors: (Constant), Mean b. Dependent Variable: MyAccess

Table 4 analyzes the statistical significance of the regression models. The findings from model 1 (F (1, 28) = 477.48, p < .05) demonstrate that machine scores are significant predictors of human rating scores for specified topics. Furthermore, the results from model 2 (F (1, 28) = 8.125, p < .05) similarly show that machine scores significantly predict human rating scores for unspecified topics.

Table 4

Regression: ANOVA Results

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.190	1	13.190	477.482	.000
	Residual	.773	28	.028		
	Total	13.963	29			
2	Regression	3.745	1	3.745	8.125	.008
	Residual	12.907	28	.461		
	Total	16.652	29			

Model 1: Defined Topics: a. Dependent Variable: MyAccess; b. Predictors: (Constant), Mean
Model 2: Undefined Topics: a. Dependent Variable: MyAccess; b. Predictors: (Constant), Mean

To assess the difference in predictive capability between machine ratings and human ratings across two contexts (defined and undefined topics), a z-test was conducted. The findings (Table 5) indicated that the predictive power of machine scores for defined topics is significantly greater than that for undefined topics (z = 6.025, p = .000).

Table 5
Comparisons of Human Raters' Prediction Powers of Defined and Undefined Topics of Machine Scores

		Machine's Scores	
		Z	**P**
Defined Topic	Undefined Topic	6.025	.000

To address question 3, three one-sample t-tests were run (Table 6) on the data obtained from the questionnaire.

Table 6
One-Sample T-Test: Accuracy, Autonomy, and Interaction

	Test Value = 3.5					
					95% Confidence Interval of the Difference	
	t	**Df**	**Sig. (2-tailed)**	**Mean Difference**	**Lower**	**Upper**
Accuracy	4.604	29	.000	.43333	.2408	.6258
Autonomy	2.858	29	.008	.22333	.0635	.3832
Interaction	-.348	29	.731	-.03000	-.2066	.1466

The third question centered on the enhancement of accuracy, autonomy, and interaction, leading to the establishment of an anticipated mean value of 3.5 (the standard mean of 3 plus 0.5). The test results indicated that, according to students' self-assessments, MyAccess Home Edition significantly enhances students' accuracy (t(29) = 4.604, p = .000 < .01) and autonomy (t (29) = 2.858, p = .008 < .01); however, it does not improve students' interaction (t (29) = -0.348, p = .731 > .05).

6. Discussion

This study rigorously investigated the validity argument of an essay-scoring machine known as MyAccess Home Edition, utilizing the Evaluation framework established by Williamson et al. (2012). To address the research questions concerning the correlation between human rater scores and machine-generated scores, a Pearson Product-Moment correlation analysis was conducted. The results indicated a strong correlation in scoring both defined and undefined essay topics, with a notable alignment between the scores assigned by the machine and those given by human raters. This is noteworthy for Iranian educators, as it may enhance their confidence in utilizing this machine for evaluating students' essays, particularly in preparatory courses for exams such as IELTS and TOEFL. Furthermore, the outcomes of the first research question corroborated the findings of Hoang and Kunnan (2016) regarding MyAccess; however, the results

diverged from their study, which reported a lack of correlation between the undefined essay scores from MyAccess and human scores. Although our results showed that moderate correlation for undefined topics, the correlation was significant, accompanied by a medium effect size.

To answer the second research question, the predictive power of machine scores over human raters' scores was explored. The results of this part of the study were unique and were never explored by any previous study on MyAccess. Based on this part, machine scores can predict 94.5 percent of human scores in defined topics and 22.5 percent for undefined topics. This significant predictive capability for defined topics indicates the machine's reliability in structured environments. However, the lower predictability for undefined topics suggests a need for ongoing development to handle less structured prompts effectively. These results emphasize the machine's current strengths and areas for improvement, providing a roadmap for future advancements in automated essay scoring.

The study also investigated the perceptions of EFL learners regarding MyAccess Home Edition, with a particular emphasis on the dimensions of accuracy, autonomy, and interaction. The results indicated that students viewed the tool as more advantageous for improving their accuracy and autonomy rather than for enhancing interaction. These findings are consistent with the research conducted by Wang et al. (2013), highlighting the importance of structured writing guidance in developing autonomous writing skills. In terms of autonomy, students expressed a favorable attitude, likely due to the writing samples and processes provided by the machine, which familiarized them with composing various types of paragraphs, including narrative and discursive forms, as well as different essay types, such as advantage-disadvantage essays. The guidelines offered by the machine-assisted learners are intended to equip students with the skills necessary for independent writing after their experience with the tool. This outcome is corroborated by a study conducted by MacArthur et al. (2016), which revealed that students who prioritized conventions and mechanics tended to be less proficient writers compared to those who recognized the importance of content and structure. This suggests that students who emphasize accuracy (structure) are more likely to develop writing skills than those who become overly reliant on writing evaluation systems.

Regarding the third research question, 15 students who participated in the study were asked to respond to 9 questions, constituting the qualitative component of the research to provide a mixed-methods approach. The findings from this segment validated the quantitative results related to research question four, which focused on students' self-reports concerning accuracy, autonomy, and interaction. Overall, MyAccess Home Edition demonstrated its reliability as a tool for evaluating students' essays, fostering their development as autonomous writers, and enhancing their accuracy. Nevertheless, there is a need for a more comprehensive user guide and training to improve the interaction between students and the machine. Providing a more interactive and engaging user experience could further support students' writing development, making the tool more comprehensive and effective.

In line with this research, Ramesh and Sanampudi (2022) conducted a systematic literature review of automated essay scoring systems, highlighting the challenges and limitations of current models. Their review emphasizes the importance of considering content relevance, coherence, and other parameters in essay evaluation, which aligns with the findings of our study. Similarly, Misgna et al. (2025) explored the use of deep

learning models for automated essay scoring and feedback generation, stressing the need for models that can explain the specific patterns and features used for scoring. This approach can enhance the transparency and effectiveness of automated essay scoring tools, providing valuable insights for future developments in this field.

Furthermore, the application of neural networks in automated essay scoring has been investigated in a study published in Nature Research Intelligence. This research focuses on improving the accuracy and reliability of automated scoring by considering multiple dimensions of essays, such as linguistic, semantic, and structural features. Integrating these advanced techniques into MyAccess could further enhance its scoring capabilities and provide more nuanced feedback to students. Additionally, Shermis and Hamner (2013) concluded that while many systems show strong performance, continuous advancements and regular updates are necessary to maintain their reliability and relevance. This emphasizes the need for ongoing development and refinement of MyAccess to keep it aligned with evolving educational standards and practices. In the context of student perceptions and engagement, Dikli and Bleyle (2014) found that timely and detailed feedback from automated systems can significantly boost student motivation and engagement. This aligns with our findings on the perceived benefits of MyAccess in enhancing accuracy and autonomy, further supporting its integration into educational settings.

7. Conclusion and Pedagogical Implications

The findings of this research indicate a notable and positive relationship between the evaluations provided by humans and those generated by machines. While the Z test demonstrated a significant disparity in the predictive capabilities of machine scores compared to human raters for both undefined and defined topics, it was established that the machine served as a significant predictor in both scenarios. According to students' self-reports, MyAccess Home Edition was found to enhance students' accuracy and autonomy significantly, although it did not lead to a notable increase in student interaction.

This research carries important implications and contributes new perspectives to the existing literature on how Automated Writing Evaluation (AWE) systems can assist educators in saving considerable time when assessing student essays and in various other educational practices. The outcomes of this study also provide insights into the application of frameworks for human raters concerning correlations. More specifically, the results present valuable pedagogical implications for English teachers, instructors of IELTS and other high-stakes examinations, course developers, test creators, and researchers in language assessment. These findings assist IELTS educators in deepening their understanding of the characteristics and scope of AWE scoring.

The current study produced several beneficial and intriguing results regarding the implementation of MyAccess Home Edition within Iran's educational framework. However, there remain additional opportunities to investigate areas related to this study. Primarily, it is essential to utilize other research tools over an extended investigation period, particularly through longitudinal studies, to yield further insights and bolster the validity argument for MyAccess Home Edition. The justification for conducting such longitudinal studies is to achieve more precise results compared to those derived from a one-month assessment. Furthermore, it would be valuable to explore and contrast the validity argument of the same assessment tool across various examinations, such as IELTS and TOEFL, to determine the tool's accuracy in relation to different tests.

Moreover, it is essential to analyze various Automated Writing Evaluation (AWE) tools concerning a single examination, focusing not only on scoring but also on the feedback provided and the perceptions of students. The study's small sample size (30 Iranian IELTS students) and its focus on a single geographical and cultural context. Future studies with larger and more diverse samples will be essential to build on this work and enhance the generalizability of the results. Despite its limitations, this study contributes to our understanding of EFL learning in Iran and offers a starting point for more expansive research efforts. While this study focuses on the quantitative aspects of scoring, Future research could address this limitation by conducting a detailed comparative analysis of the feedback provided by AWE systems and human raters. This would not only enhance our understanding of the effectiveness of AWE feedback but also provide valuable insights into how it can be integrated with human feedback to support EFL learners' writing development.

Acknowledgments

I would like to thank the participants and raters who took part in this study for their valuable support and contributions to this study.

Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

References

Ben-Simon, B., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning, and Assessment, 6*(1), 1–47.

Bitchener, J., & Storch, N. (2016). *Written corrective feedback for L2 development*. Multilingual Matters.

Brown, H. D. (2014). *Principles of language learning and teaching* (6th ed.). Pearson Education.

Burstein, J., Beigman Klebanov, B., Elliot, N., & Molloy, H. (2016). A left turn: Automated feedback and activity generation for student writers. *Proceedings of the 3rd Language Teaching, Language & Technology Workshop* (pp. 1–10).

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Chapelle, C., & Jamieson, J. (2008). *Tips for teaching with CALL: Practical approaches to computer-assisted language learning*. Pearson Education.

Chircop, J. (2005). Student achievement soars with use of quality strategies. *Educational Leadership, 63*(3), 45–48.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (Research Report No. RR-04-04). Educational Testing Service.

Coniam, D. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCALL, 21*(3), 259–279. https://doi.org/10.1017/S0958344009000147

Cowie, N. (1995). Students of process writing need appropriate and timely feedback on their work, and in addition, training in dealing with that feedback. *Saitama University Review, 31*(1), 181–194.

Daghbandan, M. (2015). *The effect of computer-assisted programs on writing accuracy: Grammarly software* [Master's thesis, University of Tehran].

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Elliot, S. (2003). IntelliMetric: From here to validity. In J. C. Burstein (Ed.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Lawrence Erlbaum.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*(2), 111–127. http://dx.doi.org/10.1076/1049-4820(200008)8:2;1-B;FT111

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2–3), 285–307. https://doi.org/10.1080/01638539809545029

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2), 939–944.

Graham, S. (2020). The sciences of reading and writing must become more fully integrated. *Reading Research Quarterly, 55*(3), 35–44. https://doi.org/10.1002/rrq.332

Graham, S., Banales, G., Ahumada, S., Muñoz, P., Alvarez, P., & Harris, K. R. (2020). Writing strategies interventions. In K. Newton (Ed.), *Handbook of strategies and strategic processing* (pp. 141–158). Routledge.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly, 13*(4), 359–376. https://doi.org/10.1080/15434303.2016.1230121

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, *44*, 100450. https://doi.org/10.1016/J.ASW.2020.100450

Knight, S., & Buckingham Shum, S. (2017). Theory and learning analytics. In C. Lang, G. Siemens, A. F. Wise, & D. Gašević (Eds.), *The Handbook of Learning Analytics* (pp. 17–22). Society for Learning Analytics Research.

Landauer, T. K., Laham, D., & Foltz, P. W. (2001). Automated essay scoring. *IEEE Intelligent Systems, 16*(5), 27–31.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*(1), 92–108. https://doi.org/10.1080/10888438.2011.536130

Li, M. (2005). *A tentative study of e-grading mechanisms in the teaching of writing* [Master's thesis, National Tsinghua University].

MacArthur, C. A., Philippakos, Z. A., & Graham, S. (2016). A multicomponent measure of writing motivation with basic college writers. *Learning Disability Quarterly, 39*(1), 31–43. https://doi.org/10.1177/0731948715583115

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives*. Corwin Press.

McNamara, D. S., & Allen, L. K. (2018). Toward an integrated perspective of writing as a discourse process. In M. Schober, A. Britt, & D. N. Rapp (Eds.), *Handbook of Discourse Processes* (2nd ed., pp. 362–389). Routledge.

Mehrani, M. B. (2017). A narrative study of Iranian EFL teachers' experiences of doing action research. *Iranian Journal of Language Teaching Research, 5*(1), 93–112.

Misgna, H., On, B. W., Lee, I., & Choi, G. S. (2025). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review, 58*(2), 1–40. https://doi.org/10.1007/s10462-024-11017-5

Qassemzadeh, A., & Soleimani, H. (2016). The impact of feedback provision by Grammarly software and teachers on learning passive structures by Iranian EFL learners. *Theory and Practice in Language Studies, 6*(9), 1884–1894. https://doi.org/10.17507/tpls.0609.23

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Reeves, D. B. (2007). The principal and proficiency: The essential leadership role in improving student achievement. *Leadership Compass, 4*(3), 1–3.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 313–346). Routledge.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education, 26*(3), 247–259. https://doi.org/10.1080/02602930120052404

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005

Suvorov, R., & Hegelheimer, V. (2013). Computer-assisted language testing. *The Companion to Language Assessment, 2*(2), 594–613. https://doi.org/10.1002/9781118411360.wbcla083

Tafazoli, D., Nosratzadeh, H., & Hosseini, N. (2014). Computer-mediated corrective feedback in ESP courses: Reducing grammatical errors via e-mail. *Procedia - Social and Behavioral Sciences, 136*, 355–359. https://doi.org/10.1016/j.sbspro.2014.05.341

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26*(3), 234–257. https://doi.org/10.1080/09588221.2012.655300

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157–180. https://doi.org/10.1191/1362168806lr190oa

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335–353. https://doi.org/10.1177/0265532210364643

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291–300. https://doi.org/10.1177/0265532210364643

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391–412. https://doi.org/10.1207/S15324818AME1504_04