Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327
مجلة اللغة الانكليزية في العالم العربي

Arab World English Journal (AWEJ) Volume 9. Number 2. June 2018 Pp. 157 -174
DOI: https://dx.doi.org/10.24093/awej/vol9no2.11

Human versus Automated Essay Scoring: A Critical Review

Beata Lewis Sevcikova
Applied Linguistics Department, College of Humanities
Prince Sultan University, Riyadh, Saudi Arabia

Abstract
In the last 30 years, numerous scholars have described the possible changes in marking writing assignments. The paper reflects these developments as it charts the paths recently taken in the field, evaluates automated and human essay scoring systems in academic environments and analyzes the implications that both systems offer. In recent years, ways and opportunities for giving feedback have changed as computer programs have been more widely used in assessing students writing. Numerous researchers have studied computerized feedback and its potential. Different problems, such as quality of this type of feedback, validity, and reliability have been analyzed. This critical review examines two major types of academic writing support. The objective of the study based on the literature review is to examine the potential support of human and automated proofreaders for teaching and learning purposes.
Keywords: assessment, rubrics, feedback, writing, automated essay scoring, human raters

Cite as: Lewis Sevcikova, B. (2018). Human versus Automated Essay Scoring: A Critical Review. *Arab World English Journal, 9* (2). DOI: https://dx.doi.org/10.24093/awej/vol9no2.11

Introduction

The movement toward the more frequent use of computers in assessing writing has been coupled with an assumption that computer-based feedback should be more efficient and accurate than traditional methods. Computer-based program innovators aimed to create more authentic assessments of tasks and constructs. All these innovations have been included in computer-based assessment tools. Computerized scoring of essays, particularly relevant to large-scale language testing programs and technological advances lead to improved understanding needed to support learning progress.

Advanced writing skills are an imperative aspect of academic performance. However, students rarely achieve advanced scores on assessments of writing skills. To attain higher levels of writing performance is an achievement "through deliberate practice that trains writers to develop executive control through repeated opportunities to write and through timely and relevant feedback." (Kellogg & Raulerson, 2007, p. 237). Based on their study, automated essay scoring software may offer a way to ease the requirements for marking and assessing writing performance. It also can substantially increase the amount of writing practice that students receive. “Regardless of whether an assessment is scored by human raters or by some automated mechanism, a goal of any assessment is to ensure that the construct is appropriately represented in the final scores or outcomes of the assessment” (Williamson, Mislevy & Bejar, 2006, p. 76).

Different writing tests and assignments are designed to score the written material of students differently and to identify whether the students have the writing skills as per the requirements. Nowadays, two major types of scoring systems are frequently used: Human Essay Scoring and Automated Essay Scoring. Both have advantages and disadvantages (Bridgeman, Trapani & Attali, 2012).

Zhang (2013) states, “Essay scoring has traditionally relied on human raters, who understand both the content and the quality of writing” (p. 1). From this perspective, an essay is evaluated by the evidence that provides the following abilities from a good writer (Bridgeman, Trapani, & Attali, 2012):

- A clear statement of the perspective of the issue to be discussed and the analysis of the relationship between the writer’s perspective and the perspective or perspectives of others.
- Development of an idea or ideas.
- Supporting the developed idea or ideas with reasons and examples.
- Organization of the idea or ideas logically and clearly.
- Communicating the idea or ideas efficiently using standardized written English.

Nevertheless, the increased use of the constructed response items, as well as the increased number of students, has made the important question of relating to the viability of the human scoring alone. According to Zhang (2013), the scoring method where only human effort is involved is lengthy, expensive and requires a lot of logistical effort. Furthermore, this process of scoring usually depends upon the judgment of a less-than-perfect human (Shermis & Burstein, 2003).

Talking about the efficiency in scoring students' written work, using computers to evaluate and assess students' assignments, make the process more effective and less expensive. The objective of the current research is to analyze both scoring systems and their advantages and disadvantages. Additionally, it will review the opinions of the researchers and scholars favoring either option from their perspectives, using logical arguments. Specifically, the current study aims at investigating the usefulness and the validity of automated essay scoring versus human scoring. The debate surrounding this comparison can also be found in academia, media and the public. Researchers are most concerned about the use of automated essays scoring within the standardized tests, as well as within the context of electronic learning environments, primarily used inside and outside of classrooms. Important things for the test developers, educators, and policymakers are to have adequate knowledge relating to the strengths and weaknesses of both scoring methods so that the prevention of misuse can be attained (Attali & Burstein, 2006). From this perspective, the present work contrasts the distinguishing features of the two scoring methods by elucidating their differences and discussing their practical implications for testing and learning purposes.

Writing in the Academic Environment

In contemporary research, authors draw attention to the 'problem' of student writing in higher education. This problem is described from various perspectives such as from the "perspective of 'non-traditional' student-writers as they attempt to involve in academic writing or from the perspective of a cultural-historical tradition of scientific rationality" (Lillis & Turner, 2001, p. 57). Researchers argue that academic writing practices need to reach a complex understanding of what is involved in student writing. Some have expressed concerns that English for Academic Purposes (EAP) is too focused on the needs of L2 (second language) and therefore, fails to make an impact on "mainstream writing instruction" (Wingate & Tribble, 2012, p. 481).

Learning writing skills is a long process. If a writer accomplishes it, he/she becomes expert in the complex cognitive domains. Academic writing at advanced level challenges not only student language skills but also thinking, memory and general knowledge (see Figure 1). In theory, writers can use everything they have learned and remembered. Their cognitive systems for memory can only be used if they can retrieve knowledge from long or short-term memory. It means that writing is closely connected to thinking. Based on the various studies, good writers are often good thinkers and problem solvers. What to write and how to say it is a decision-making process which reflects an individual's knowledge and cognitive skills.

When correcting writing assignments or tests, educators and testers usually focus on audience, purpose, organization and style, while taking into consideration the lexical and grammatical means by which a formal written style is achieved. The writing critiques provided by educators is daunting for many L2 writers. Students learn how to gain awareness of stylistic and generic conventions, and to practice different aspects of academic writing. One dilemma that arises here is that many students with a reasonable English proficiency level require more grammatical and lexical back-up, whereas students who already possess a good command of the written language do not have either time or patience to work systematically (Swales & Feak, 2004).

Regarding the doubts mentioned above, it is apparent L2 writers should be empowered to use language efficiently in real-world conditions and avoid simplistic "recipes" for writing. Educators assume that students develop over time, in response to education, practice, and feedback. Their assignments or tests are analyzed for overall quality, grammatical accuracy and syntactic complexity using different assessment styles. This pedagogical practice helps students to improve at the discourse level, linguistic complexity, and language accuracy. A grading rubric aids this pedagogical practice because it sets a clear set of criteria used for evaluating a specific type of work or performance. It provides detailed guidelines for a marker and helps to improve objectivity. Typically, a grading rubric includes specific criteria, levels of performance such as 'excellent, poor, fair,' scores, and clear descriptors of the performance. This tool then serves as a marker for assessing objectively an assignment. Grading rubrics represent effective and efficient tools which can help clarify an educator's expectations and help students to meet them. Rubrics also encourage students to improve their writing assignments, ensure consistency in grading and serve as proper evidence of student performance.

Even though rubrics make expectations and criteria for writing explicit, promote learning and/or improve instructions and facilitates feedback and self-assessment, there is still a lack of information in the literature describing the actual effectiveness of the rubric as an assessment tool in the hands of the students (Jonsson & Svingby, 2007; Hafner & Hafner, 2003). According to recent findings, raters' biases have related to gender, language command, and for some raters, mechanical characteristics of students' writing are more important than the content even when they used a rubric (see Appendix A). Using rubrics may not improve the reliability or validity of assessment if raters are not well trained on how to use them effectively. However, in general, rubrics lead to a more reliable and less biased assessment (Rezaei & Lovorn, 2010).

Human Versus Automated Essay Scoring Systems

Human Scoring

Several large-scale testing programs or summative tests in academia include at least one essay writing task. Human raters typically assess an essay's quality according to a scoring rubric that states the criteria an essay to meet a specific score level. Zhang (2013) argues that the quality of an essay is gauged by human raters with the help of a scoring rubric identifying the set characteristics (see Table 1). In other words, the human essay scoring system is aligned with a certain score level that is known as merit. A powerful learning aid in writing is feedback given by an instructor. Assessing written texts is probably the most challenging task for an instructor because it requires practice, routine and understanding of marking guidelines (even though assessing an individual's writing is not a routine task at all). It is well accepted that holistic grading is faster than an analytic one. Analytic evaluation focuses on different features of the text, such as mechanics, coherence, and content. Of course, even holistic grading can be extremely time-consuming. In some studies, instructors claim that they do not assign writing a research term paper longer than 5,000 words because it takes too long to mark papers.

Although human essay scoring systems are time-consuming and require much effort, there are certain strengths in these systems. For example, the information given in the text is sent through a

cognitive process and thus has a connection with prior knowledge. Human scoring is also based on the comprehending of the given content which is the reason why human raters can make a judgment on the quality of the text (see Appendix B). Zhang (2013) states, "Trained human raters are able to recognize and appreciate a writer's creativity and style (e.g., artistic, ironic, rhetorical), as well as evaluate the relevance of an essay's content to the prompt" (p. 2). Likewise, a human rater can evaluate the critical thinking skill of an examinee, including the factual accuracy of the claims as well as the quality of argumentation presented in the essay.

Despite strengths, there are limitations such as human raters' specialized training requirements, experience in assessing writing, individual variations in understanding, interpretation and implementation of a rubric. Moreover, raters must be educated in how to understand and apply a scoring rubric which can be costly. All these issues are reflected in rating competencies which are not always consistent and bias-free. It is not always possible to avoid all the problems mentioned as mentioned earlier, especially if many essays are supposed to be assessed. According to Williamson at al. (2010), the process of essay scoring in which human effort is used is also time-consuming and thus if a considerable quantity is required to be scored, it becomes cumbersome. The other notable disadvantage of human essay scoring includes its limitations of consistency and objectivity.

In the light of the reviewed literature, it can be concluded that humans possess the ability to make holistic conclusions under the impact of numerous interacting factors. Understanding weaknesses of human raters can help to minimize disagreements and accept individual perspectives of each marker.

Automated Essay Scoring

Automated essay scoring systems provide fine-tuned and instantaneous feedback that is helpful in practicing and improving writing skills simultaneously. One of the unique strengths of the automated essay scoring system is its efficiency and consistency. Since the automated essay scoring system is based on a computer application, it is not influenced by any external factors such as deadlines, nor is it attached emotionally to any piece of work. According to some researchers, there is no bias, preconceptions or stereotypes in a computer-based application. From this perspective, the automated essay scoring system has the potential to achieve greater objectivity when compared to a human essay scoring system.

Dikli and Bleyle (2014) argue that the increased reliability of the automated essay scoring system has led to increased demand for this type of system. The recent development of automated essay scoring can, therefore, boost the number of writing assignments. Shermis and Burstein (2003) examined various computerized feedback, and scoring methods originated from cognitive psychology and computational linguistics. Based on their research, students need more opportunities to write as they can profit from immediate computer-based feedback. This type of feedback can also motivate them and help to improve their results prior to submitting their assignments for assessment.

Incorporating large numbers of writing assignments challenges instructors. The effort to evaluate writing assignments is inevitable for the educational process to support educational achievements. An essay or open-ended question-based testing encourages students' critical thinking and a deeper level of knowledge. Thus, grading and feedback on written texts are essential not only as an assessment tool but also as a feedback device supporting students' learning, thinking and writing.

According to Foltz, Laham and Landauer (1999), "essays have been neglected in many computer-based assessment applications since there exist few techniques to score essays directly from a computer" ("Introduction," para 1). In their study, a statistical analysis of essays scored by computer programs proves the accuracy of the computer-based feedback. The analysis of the text is based on Latent Semantic Analysis (LSA). They prove that LSA can capture semantic similarity of words as well as recognize the coherence of texts on readers' comprehension. Based on numerous statistical analysis and 'text training,' LSA scored as well as average test-takers.

There are several factors involved in evaluating a writing assignment, from mechanical features, such as grammar, spelling, and punctuation, to abstract features, correctness,
the fluency, elegance, and comprehensibility. Evaluating all these features is not difficult but evaluating content, argument, comprehensibility, and aesthetic style are more problematic as each influence the other because each depends on the choice of words (Foltz at al., 1999).
The first attempts of computational scoring focused primarily on measures of style (Page, 1994) while LSA methods focus on the conceptual content and the knowledge communicated in an essay as LSA was trained on domain-representative texts (such as textbooks, samples of writing, journal articles).

As Foltz at al., (1999) explains, "several techniques have been developed for assessing essays. One technique is to compare essays to ones that have been previously graded. A score for each essay is determined by comparing the essay against all previously graded essays" ("Evaluating the effectiveness of automated scoring," para 1). This holistic scoring method evaluates the overall similarity of content, actually, it determines how adequately the overall meaning resembles that of previously graded essays. This approach is similar to the holistic scoring approach used by human markers. The research also judges LSA's reliability which is equivalent to human raters' reliability.

Dikli (2006) points out, "Automated Essay Scoring (AES) is defined as the computer technology that evaluates and scores the written prose. AES systems are mainly used to overcome time, cost, reliability, and generalizability issues in writing assessment" (p. 3). On the other hand, Williamson at al. (2010) argues that human scoring is not the only option today, where technology is available everywhere, to score the constructed-response (CR) items. However, much practical experience and years of research demonstrate many challenges.

, Bernstein, Foltz, and DeLand (2011) argue that automated essay scoring provides and promotes consistency in location and time by enabling a precise trend analysis, as it offers

comparable feedback for use at school, classroom, state, or district level. The rapid growth of automated essay scoring can be observed on a significant scale and is probably due to the reason that the system has the potential capability to produce the scores quicker and more reliably. Be that as it may, it is considerably costlier (Topol, Olson & Roeber, 2010). On the other hand, Zhang (2013) points out that the noticeable shortcomings to be found in the human essay scoring system can be eliminated by using the automated essay scoring systems available online. The state-of-the-art systems of today involve the construct-relevant combination of quantifiable text features for computer-based scoring to measure the quality of a written essay. Deane, (2013) claims that AES systems exhibit clear similarities with overall performance and can adequately distinguish between students and apply a broader writing construct from those for whom text production constitutes a significant barrier to achievement. Nevertheless, an automated essay scoring system works solely with variables that are to be extracted as well as combined mathematically (Roscoe, Crossley, Snow, Varner & McNamara, 2014).

, Zhang (2013) also adds to the discussion by favoring the automated essay scoring system, stating that it has the potential capability to assess the essays across grade levels. The improvement of automated writing scoring tools has made it possible to complement lecturer input with immediate scoring and qualitative feedback to inform students' writing development. Zhang's (2013) findings added to previous work on the usefulness of such programs and justified the use or non-use of automatically generated scores for classroom-based formative assessment (Li, Link, Ma, Yang & Hegelheimer, 2014).

Scoring with Professional and Online Automated Scoring Applications

There are two types of AES systems available for individual as well as classroom applications:

1. Professional Scoring Applications
2. Online Scoring Applications

Professional Automated Scoring Applications

These applications were designed directly for classroom implementation. School authorities in different countries use standardized tests for different subjects to evaluate individually and collectively to ensure the quality of language education in various educational settings. One of the most common and most frequently used components is essay writing. In some countries, educators evaluate the essays of their students and determine their scores accordingly, while in other countries, including Canada and United States, educators use two blind raters to evaluate the essays of their students (Burstein, Tetreault & Madnani, 2013). These blind raters involve the scoring process in the standardized tests. However, when both blind raters agree on the gained score, the score is defined as satisfactory. On the other hand, when both blind raters do not agree on the gained score, a third rater is used to help with the final decision (Smolentzov, 2013).

Testing Service (ETS), the most significant educational assessment and testing organization, has been using the professional computer application for the last two decades. ETS research (2017) concentrated on its holistic scoring features. The professional application

developers' driving concept focused on the same features that human raters focus on. From the beginning, they used the scoring guide for the human raters, and the priority was not to measure essay length. Based on the Burstein and Chodorow (1999) study,

> The features currently used by the system are syntactic features, discourse cue words, terms and structures, and topical analysis, specifically, vocabulary usage at the level of the essay (big bag of words) and at the level of the argument. An argument, in this case, generally refers to the different discussion points made by the writer. (p. 69)

The criteria used by ETS for the evaluation of the writing skills is a web-based service using ETS providing an instant score reporting, as well as diagnostic feedback. This instructor-led, web-based application is used to help students plan, write, and (if required) revise their writing. The main purpose of the application is to give the students instant diagnostic feedback as well as the opportunity to practice the writing skills at their own pace. Students use this feedback to evaluate their writing skills and thus can identify the areas they need to improve (Ramineni & Williamson, 2013). A significant advantage of using this application includes the ability for the students to develop written skills independently, as they receive automated and constructive feedback.

According to ETS Research (2017), the specific features of the application include:

- mechanics such as capitalization
- usage of prepositions
- grammatical errors such as subject-verb agreement
- discourse structure, the thesis statement or the main points
- style, for example, word repetition
- sentence variety
- vocabulary usage, the relative complexity of vocabulary
- discourse quality
- source use

Murray and Orii (2012) point out that standardized test should not be evaluated by using 'manual effort' to score the written work of students. This is because the latest technology has advanced machine learning methods that can not only save time and effort but also reduce the chances of errors. Their paper also compares the two versions of the specialized computer applications. It focuses on the specific features such as mechanics, discourse structure, the thesis statement or the main points, style, namely word repetition, sentence variety, vocabulary usage, the relative complexity of vocabulary, discourse consistency quality, and source use.

One of the new features included in the new version of the application used by ETS is the qualitative feedback. According to Burstein and Wolska (2003), the output focuses on the feedback errors such as grammatical, sentence structure and usage, and other mechanics. This feedback also includes the comments related to the writing styles. Some of the new features of this application include standardized checking of the length of an essay submitted and altering a definition to take account of non-monotonic relationship along with the human score. There are some other distinguishing features that make this version generate standardized scores (Ramineni & Williamson, 2013).

Burstein, Marcu, and Knight (2003) believe the application as mentioned above automatically identifies the sentences used in an essay that correspond to essay-discourse categories. The application uses natural language processing such as the background of the topic, statement of the thesis, presentation of the main idea, and the conclusion methods. The overall score is calculated considering the individual items including the score of the thesis, the main points, the supporting ideas, and the elements used in the conclusion of the essay. Some of the other features are lexical complexity, usage of prompt-specific vocabulary, and essay length.

Attali and Burstein (2006) conclude, the application "…uses a small and fixed set of features that are also meaningfully related to human rubrics for scoring essays" (p. 19). Their study shows that the advantages of the new features integrated into the latest version can be utilized to generate the automated essay scores and thus, are considered as standardized across the different stimuli without losing the performance. This happens because the features of the new version have higher agreement rates than that of the human scores.

Moreover, Quinlan, Higgins, and Wolff (2009) hold the opinion which states that the scoring done by the described application contains a broad set of measures having a proven ability to predict human holistic scores. Since this new version has its 'features,' large sets of scores can be aggregated into a small set of readily recognizable categories. They also conclude that the prediction of human scores can be considered as one type of score validity, whereas the construct validity can be taken separately. For example, if the essay length is to be considered, it would not be wrong to state that predictors may have lesser or greater construct relevance in modeling human holistic scores (Weigle, 2013).

In the light of the secondary data, it can be concluded that the new version of the Automated Essay Scoring application used by ETS contains more human-based holistic scores as compared to the previous version. The primary purpose of the tool is to give students instant diagnostic feedback as well as the opportunity to practice the writing skills at their own pace. Studies by many researchers conclude that the advantages of new features integrated into the new version of this professional Automated Essay Scoring application, can be utilized to generate automated essay scores and thus be considered as standardized across the different stimuli.

Online Automated Scoring Applications

Applications which were designed to be accessed by the public have two versions, free or paid. Both types help people write and communicate more efficiently. They do not require specific training to use. These are useful tools as they quickly assist in checking for various mistakes. The most advanced ones (paid applications) have been complementing teaching and learning environments since their introduction to the market. It is a fact that proofreading of one's writing can be a very demanding task. Most students who write for academic success understand that proofreading before submitting the task is essential. With proofreading, every help is welcome. The autocorrect and spell-check tools most word processors have been useful tools, but they are basic. Some paid applications can fix numerous types of errors, and they provide plenty of other features that may help students improve their grammar, vocabulary style and sentence structure

problems such as word order. They also provide detailed information about each error which can serve as a study aid:

- The style checkers detect wordiness and redundancies.
- The vocabulary enhancement tool offers synonyms and suggestions for word use.
- These applications also check for contextual spelling, grammar and punctuation inconsistencies, comma splices, the subject-verb disagreement, and they suggest proper corrections.
- Some also offer adjustments of the genre-specific writing styles, a plagiarism checker, and a vocabulary enhancement tool.
- They identify possible solutions and explanations for mistakes.

Impact on Learning

Smolentzov (2013) discusses, "Essay scores may be used for very different purposes. In some situations, they are used to provide feedback for writing training in the classroom. In other situations, they are used as one criterion for passing/failing a course or admission to higher education" (p.1). The current research emphases how rubrics can assist learners to learn, think and meet the writing requirements. Little research on AES feedback on learning has been undertaken. In general, rubrics have been designed to support and evaluate student learning. Rubrics are written in language that students can understand, they define and describe the quality of work, they point out common weaknesses in students' work and indicate how such weaknesses can be avoided. Rubrics make assessing student work faster and more efficient. "At their very best, rubrics are also teaching tools that support student learning and the development of sophisticated thinking skills. When used correctly, they serve the purposes of learning as well as of evaluation and accountability" (Andrade, 2000, p. 13).

Present days studies analyze the use of computer-assisted grading rubrics compared to other grading methods based on the efficiency and effectiveness of different grading processes for tests and writing assignments. Studies recommend using this system as it is much faster than traditional hand-grading systems. Moreover, researchers also advise using computer-assisted grading applications as they "did not negatively affect student attitudes concerning the helpfulness of their feedback, their satisfaction with the speed with which they received their feedback, or their satisfaction with the method by which they received feedback" (Anglin, Anglin, Schumann, & Kaliski, 2008, p. 51). Finally, many researchers proposed implementing AES as complementary methods of proofreading, learning and scoring to improve: (a) writing classroom practices, (b) student learning autonomy – using accessible applications on a regular basis can help students to improve bad writing habits and (c) avoid plagiarism (learning how to paraphrase, summarize and cite).

Conclusion

The review of the presented study reveals that it is important how we introduce technology into the educational environment. Utilizing technology can ease the burden human markers are under, however, student-instructor interaction can never be entirely replaced.

Offering immediate analytical feedback is nearly impossible for human raters, as it is difficult in large classes. As mentioned in the reviewed literature, with the aid of automated essay scoring systems, it is quite easy to evaluate essays through intelligent computer programs. On the other hand, human raters are most often trained to give their feedback focusing on a specific grade range that is associated with a specific set of tasks and a specific rubric.

To summarize, automated essay scoring systems (either professional or online applications), when developed or used carefully, can contribute to the efficient delivery of essay scores. It can be an important aid in the improvement of educational writing skills. As this critical review indicates, there are some logistical obstacles to incorporating technology. If we think about technology-based teaching writing or assessment in the educational environment, schools should have access to reliable and affordable essay scoring systems. Moreover, well-trained educators who understand how to use such programs is an imperative requirement. Technology-based assessment has great potential, however some of the above challenges may explain why it is not yet widely used. Alternatively, some easy accessible online applications can help to take the teaching of writing into the 21st century. Can they replace a human proofreader? The short answer is no. They still overlook some mistakes; they still do not always provide the context or feedback that a human proofreader offers.

Incorporating more technology into assessments has the potential to make teaching, learning and testing a less cumbersome task for educators and free them for more face to face interaction with the student. Online applications, instructor feedback and human markers all have something to contribute to the enhanced efficiency of language instruction.

Acknowledgements
The researcher thanks Prince Sultan University and the research group (Language Learning and Teaching Group RG-CH-2016/11/11) for supporting this research project.

About the Author:
Dr. Beata Lewis Sevcikova is an experienced educator who has participated in numerous in-house and international workshops and symposiums. In her research, she focuses on teachers' experiences in the use of new technology in the classroom. ORCID ID: 0000-0002-4347-0489

References
Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational leadership, 57*(5), 13-19.

Anglin, L., Anglin, K., Schumann, P. L., & Kaliski, J. A. (2008). Improving the Efficiency and Effectiveness of Grading Through the Use of Computer-Assisted Grading Rubrics. *Decision Sciences Journal of Innovative Education, 6*(1), 51-73.

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater [R] V. 2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-22.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27-40.

Burstein, J., & Chodorow, M. (1999, June). Automated essay scoring for nonnative English speakers. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing* (pp. 68-75). Association for Computational Linguistics.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems, 18*(1), 32-39.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. *Handbook of automated essay evaluation: Current applications and new directions*, 55-67.

Burstein, J., & Wolska, M. (2003). *Toward evaluation of writing style: finding overly repetitive word use in student essays.* Paper presented at the EACL '03 European chapter of the Association for Computational Linguistics Budapest, Hungary.

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

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment, 5*(1), 1-36.

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.

ETS Research, (2017). *ETS Research: Automated Scoring of Writing Quality*. *Ets.org*. Retrieved 18 March 2017, from https://www.ets.org/research/topics/as_nlp/writing_quality/

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). *Automated essay scoring: Applications to educational technology.* Paper presented at the World Conference on Educational Media and Technology, Seattle, WA USA.

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. *Int. J. Sci. Educ., 25*(12), 1509-1528.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144.

Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic bulletin & review, 14*(2), 237-242.

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*, 66-78.

Lillis, T., & Turner, J. (2001). Student writing in higher education: contemporary confusion, traditional concerns. *Teaching in Higher Education, 6*(1), 57-68.

Murray, K. W., & Orii, N. (2012). Automatic Essay Scoring. In: Carnegie Mellon University.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of experimental education, 62*(2), 127-142.

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. *ETS Research Report Series, 2009*(1).

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

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.

Roscoe, R. D., Crossley, S. A., Snow, E. L., Varner, L. K., & McNamara, D. S. (2014). *Writing quality, knowledge, and comprehension correlates of human and automated essay scoring.* Paper presented at the 27th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014, Florida, USA.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Erlbaum: Routledge.

Smolentzov, A. (2013). *Automated Essay Scoring: Scoring Essays in Swedish* (Doctoral dissertation, Stockholm University, 2013) (pp. 1-43). Stockholm, Sweden: Stockholm University.

Streeter, L., Bernstein, J., Foltz, P., & DeLand, D. (2011). Pearson's automated scoring of writing, speaking, and mathematics. *Retrieved on June, 25*, 2013.

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (Vol. 1): University of Michigan Press Ann Arbor, MI.

Topol, B., Olson, J., & Roeber, E. (2010). The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments. *Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved August, 2*, 2010.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*(1), 85-99.

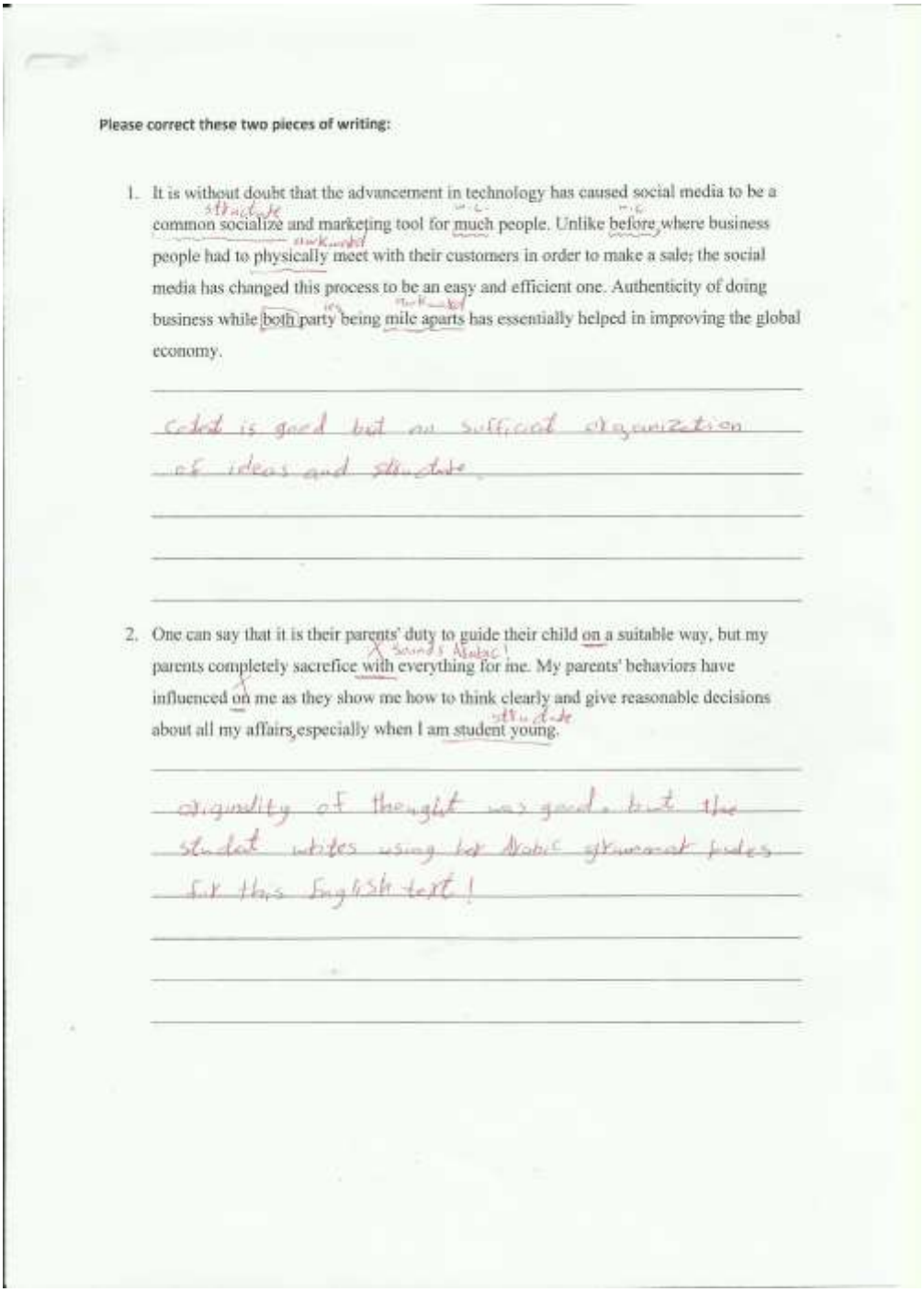
Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing (pp. 1-4370*. New Jersey, US: Lawrence Erlbaum Associates, Publishers.

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., . . . Sweeney, K. (2010). Automated scoring for the assessment of common core standards. *White Paper*.

Wingate, U., & Tribble, C. (2012). The best of both worlds? Towards an English for Academic Purposes/Academic Literacies writing pedagogy. *Studies in Higher Education, 37*(4), 481-495.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections, 21*(2), 1-11.

Appendix A. Raters' Bias Samples



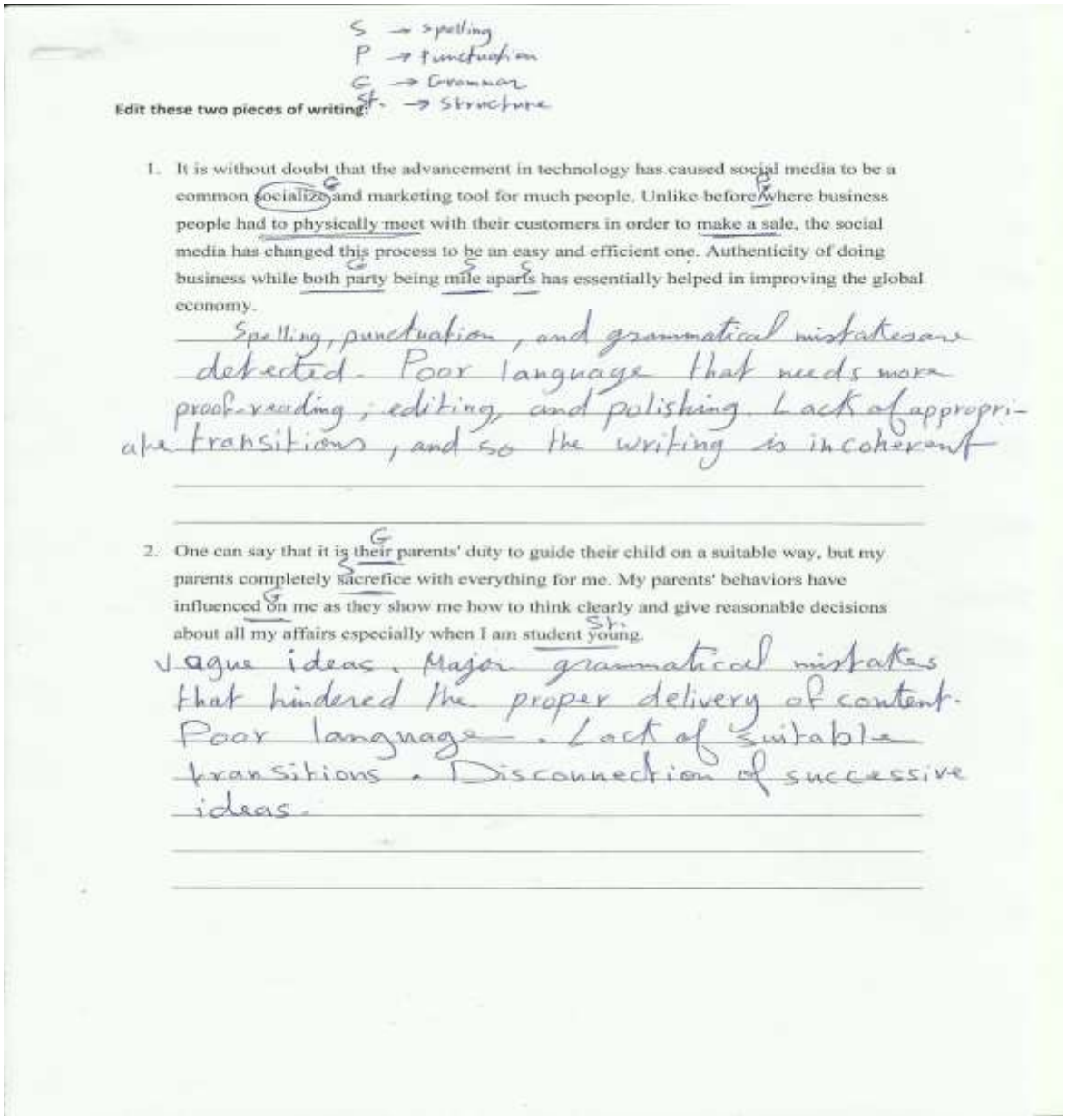
Please correct these two pieces of writing:

1. It is without doubt that the advancement in technology has caused social media to be a common socialize and marketing tool for much people. Unlike before where business people had to physically meet with their customers in order to make a sale; the social media has changed this process to be an easy and efficient one. Authenticity of doing business while both party being mile aparts has essentially helped in improving the global economy.

Coherent is good but an sufficient organization of ideas and structure.

2. One can say that it is their parents' duty to guide their child on a suitable way, but my parents completely sacrefice with everything for me. My parents' behaviors have influenced on me as they show me how to think clearly and give reasonable decisions about all my affairs especially when I am student young.

originality of thought was good. but the student writes using too basic grammar rules for this English test!



S → spelling
P → punctuation
G → Grammar
St. → Structure

Edit these two pieces of writing:

1. It is without doubt that the advancement in technology has caused social media to be a common socializ and marketing tool for much people. Unlike before where business people had to physically meet with their customers in order to make a sale, the social media has changed this process to be an easy and efficient one. Authenticity of doing business while both party being mile aparts has essentially helped in improving the global economy.

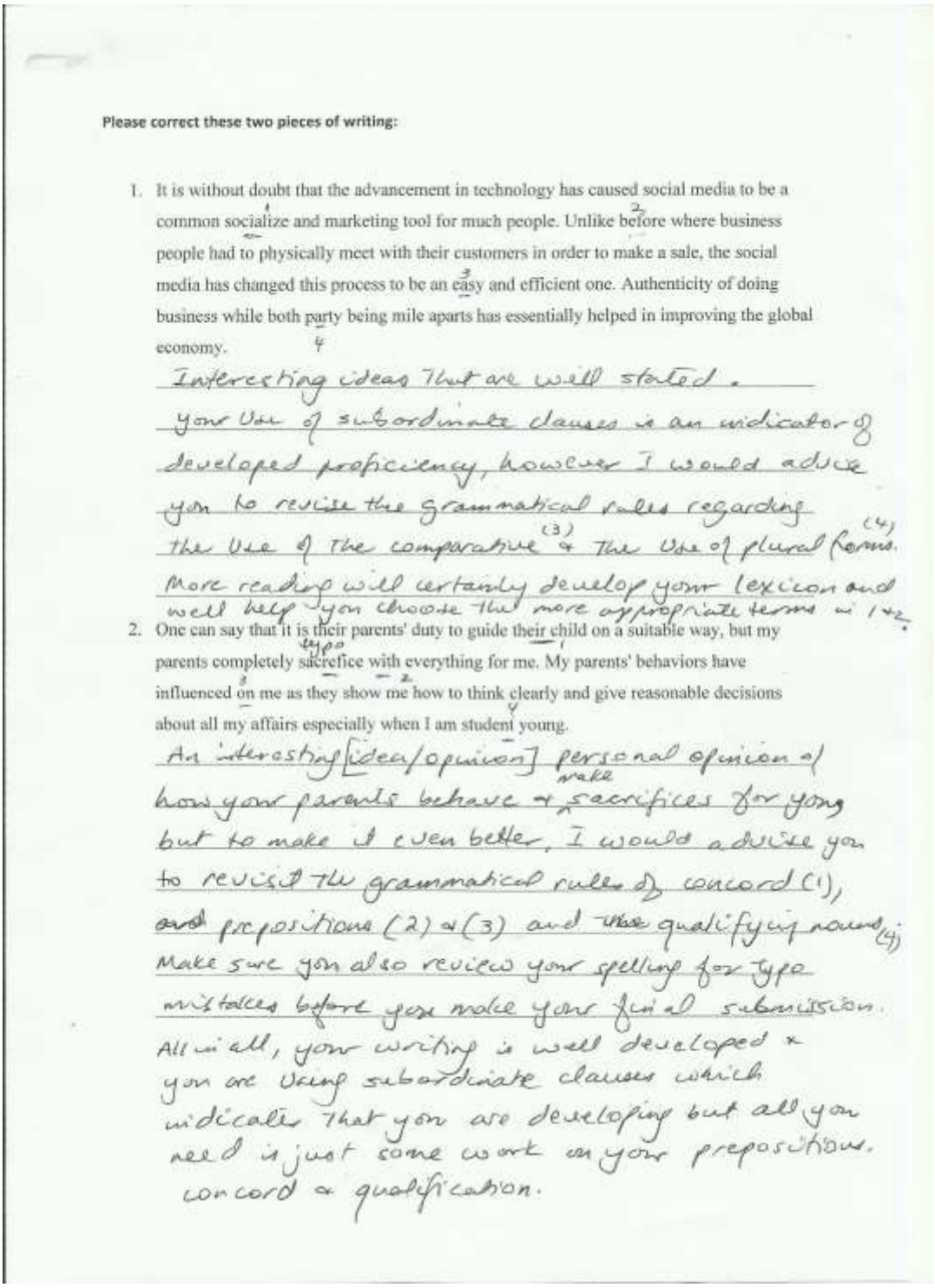
Spelling, punctuation, and grammatical mistakes are detected. Poor language that needs more proof-reading, editing, and polishing. Lack of appropriate transitions, and so the writing is incoherent

2. One can say that it is their parents' duty to guide their child on a suitable way, but my parents completely sacrefice with everything for me. My parents' behaviors have influenced on me as they show me how to think clearly and give reasonable decisions about all my affairs especially when I am student young.

Vague ideas. Major grammatical mistakes that hindered the proper delivery of content. Poor language. Lack of suitable transitions. Disconnection of successive ideas.

ISSN: 2229-9327

Appendix B. Judgments on the Quality of the Text Samples



Please correct these two pieces of writing:

1. It is without doubt that the advancement in technology has caused social media to be a common socialize and marketing tool for much people. Unlike before where business people had to physically meet with their customers in order to make a sale, the social media has changed this process to be an easy and efficient one. Authenticity of doing business while both party being mile aparts has essentially helped in improving the global economy.

Interesting ideas that are well stated. your use of subordinate clauses is an indicator of developed proficiency, however I would advise you to revise the grammatical rules regarding the use of the comparative (3) & the use of plural forms (4). More reading will certainly develop your lexicon and well help you choose the more appropriate terms in 1+2.

2. One can say that it is their parents' duty to guide their child on a suitable way, but my parents completely sacrefice (typo) with everything for me. My parents' behaviors have influenced on me as they show me how to think clearly and give reasonable decisions about all my affairs especially when I am student young.

An interesting [idea/opinion] personal opinion of how your parents behave & sacrifices (make) for you but to make it even better, I would advise you to revisit the grammatical rules of concord (1), and prepositions (2) & (3) and the qualifying nouns (4). Make sure you also review your spelling for typo mistakes before you make your final submission. All in all, your writing is well developed & you are using subordinate clauses which indicates that you are developing but all you need is just some work on your prepositions, concord & qualification.

ISSN: 2229-9327

Table 1. *Strengths and weaknesses in human and automated scoring of essays*

	Strengths	**Weaknesses**
Human Raters	Can: - Comprehend the meaning of the text being graded. - Evaluate critical thinking. - Assess creativity. - Judge the content relevance (in depth). - Evaluate logic and quality of argumentation. - Assess factual correctness of content and claims. - Judge audience awareness.	- Cannot maintain consistency while evaluating assignments. - Cannot completely avoid certain subjectivity in the evaluation. - Different standards of strictness. - After marking numerous assignments, scale shrinkage errors appear. - Stereotyping errors. - Require highly specialized training in marking and calibrating marks.
Professional Automated Scoring Applications	Can: - Detect a surface-level content. - Correct grammar and mechanics. - Evaluate organization and style (some). - Detect plagiarism. - Evaluate vocabulary and suggest an enhancement. - Assess objectively (not influenced by emotions). - Be consistent in grading. - Provide the same scoring over time. - Offer explanations of errors.	- Do not have background knowledge. - Cannot assess creativity, quality of argumentations, logic, quality of ideas and content development. - Can be costly. - Costly maintenance and upgrades.
Online Automated Scoring Applications	Can: - Evaluate based on genre (academic, business, technical, etc.).	- Do not have background knowledge.

	- Correct grammar and mechanics. - Evaluate organization and style (some). - Detect plagiarism (some). - Evaluate vocabulary and suggest an enhancement. - Assess objectively (not influenced by emotions). - Be consistent in grading. - Provide the same scoring over time. - Offer explanations of errors. - Some are free. - Do not require funds for maintenance and upgrades.	- Some free applications do not detect all mistakes in grammar and mechanics. - Cannot assess creativity, quality of argumentations, logic, quality of ideas and content development.

Figure 1. Stages of Cognitive Development in Writing Skills ((Adopted from Kellogg, 2006)