Automated Essay Scoring Feedback (AESF): An Innovative Writing Solution to the Malaysian University English Test (MUET)

Sing Yii Ng[1*], Chih How Bong[2*], Nung Kion Lee[1] and Kian Sam Hong[1]

[1]Faculty of Cognitive Sciences and Human Development,
Universiti Malaysia Sarawak,
93400 Kota Samarahan, Sarawak, Malaysia
ngsingyii@gmail.com
[2]Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak,
93400 Kota Samarahan, Sarawak, Malaysia
chbong@fit.unimas.my
[*]Corresponding Author

ABSTRACT

Recent advances in information and communication technology (ICT) infrastructure can be harnessed to support and improve the quality of teaching and learning of English writing skills especially for second language context where rule based support is necessary. Essay writing is indeed the most demanding tasks to both teachers and students. From conducting the class to the assigning of task as well as marking and providing feedback from teachers, whereas from drafting essays to final submission and resubmission of essays by students require on-going iterative cycles to facilitate improvement. However, a common scenario is that the iterative process takes too much time, thus resulting in limited practice. An innovative solution to imitate such process is via the Automated Essay Scoring Feedback (AESF). AESF is a networked tool that has the ability to score and provide feedback to students' essays instantaneously. With the speed that exceeds human ability and accuracy of a human scorer, it is hoped that AESF can increase the frequency of essay writing in the class that eventually results in improvement in students' performance. This paper aims to highlight the novelty and rationale of having AESF, its design and features as well as how this tool can be blended into the writing classroom, particularly for the Malaysian University English Test (MUET) extended essay writing.

Keywords: *automated essay scorer; paragraph scoring*

INTRODUCTION

Harnessing computational methods in essay marking is no longer a new issue and is being greatly expanded to large scale assessment, including Scholastic Aptitude Test (SAT), Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL) and Graduate Management Admission Test (GMAT) (Attali, Burstein, Russell & Hoffmann, 2006; Shermis, 2014). Automated Essay Scorers (AES) dated as early as 1966 by Professor Ellis Page with Project Essay Grade (PEG™)(Page, 2003) and thereafter, Criterion (Burstein, Chodorow & Leacock, 2004), Intelligent Essay Assessor (IEA), and IntelliMetric ("IntelliMetric® | Vantage Learning," n.d.). These systems are proven fast, exceeding human scoring, and reliable with a higher inter rater reliability as compared to the reliability of only human marking (Shermis, 2014).

In this paper, an Automated Essay Scoring Feedback (AESF) system is proposed to aid secondary school students in learning Malaysian University English Test (MUET) essay writing. The AESF system is a web-based instructional writing tool that can score and provide feedback to submitted essays instantaneously as demanded by users, targeting on students at the pre-University stage. AESF is developed based on Natural Language Processing and supervised machine learning framework where the scoring model is trained using a large collection of essays with different scores obtained from students on pre-determined topics. The system is hoped to simplify teachers' tasks and improve students writing ability by on-going and sufficient practice as they needed as suggested by various commercial AES (Mayes, 2014). The novelty lies in teacher autonomy, student autonomy, cultural sensitivity, and paragraph level grading.

BACKGROUND

Automation of essay marking employs sophisticated language processing technologies and statistical methods to analyse a wide range of text features with its corresponding values that are being internalised or learned by the system to score unknown essays (Li, Link & Hegelheimer, 2015). The automation process is generally similar with human holistic scoring, but with huge samples. Human evaluation of essays is usually based on marking

schemes that outline rubric that delineates specific expectation on essay responses. A moderation process is based on small samples that serve as bench marks and eventual agreement on marking between two or more graders to fine pitch the marking score (Attali et al., 2006).

Unlike humans, who can read and internalise the scoring rubric with their background knowledge and language processing skills, system on the other hand, requires a huge pool of data for learning and training before it can score accordingly (Dikli, 2006). Once, the system has internalise the text features, it can score as accurate as human scorers and more reliable than human, with great speed that excludes human weaknesses of being bias, inconsistent and having individual preferences (Shermis, 2014).

Currently, Malaysia lacks home-grown AES that is tailor-made for the Malaysian context, especially for marking extended English language essays. There are some local systems that only cater for short answer response with predetermined finite answer keys (Ab Aziz, Ahmad, Abdul Ghani& Mahmod, 2009). As for extended general English writing skills improvement system, this technology is not available as most research results published on the AES effectiveness with Malaysian students are based on commercially available systems like Criterion and My Access! (Li et al., 2015). A drawback in such system is that the grading may not be valid because the training model is based on essays written by native language users (L1) (Ene & Upton, 2014) while the marking criteria/scheme may not necessarily be the same as how a Malaysian teacher may grade their students' essays. Therefore, it is unfair to grade, second language users (L2) against L1 where essays may also be culturally different than the L2. Thus, if essays are not measured with the same yardstick, the scoring cannot be valid (Dikli, 2006).

Therefore, there is an urgent need for a tailor-made tool that can help score essays reliably and validly in the Malaysian school context. Besides, automated feedback is accepted by students and should be further improved to help L2 students to be more precise in using the language (Ene & Upton, 2014). AESF targets the Malaysian University English Test (MUET) for prototyping because MUET students are at a stage just before varsity. This is also a good platform to train students to use ICT for independent learning at the tertiary level as they will be required to use ICT extensively

for producing reports, assignments and thesis. If writing via computer is a must, then utilising MUET students in AESF development and usage can be more fulfilling for students who see the need to use ICT apart from being more critical and mature in providing feedback on the usage of the system. With this, the AESF prototype can be further improved and also be adjustable to other level of education in school.

The Development of AES

Project Essay Grade (PEG™) was one of the earliest automated essay scorers, devised by Ellis Page in 1966 using proxy measures to determine the grade of the essays (Page, 2003; Rudner & Gagne, 2001). The features include average word length, essay length, and the use of commas and semicolon (Rudner & Gagne, 2001). This system does not include aspects of semantic, lacking in human ability to organize and make meaningful transactions.

Subsequently, Intelligent Essay Assessor (IEA), a system which considers the semantic value of essays was introduced (Lemaire & Dessus, 2001). This is achieved using Latent Semantic Analysis (LSA) technique to assess essays. This scoring technique assumes that "there is a hidden semantic space in each text which is the accumulation of all words meaning" (Jiang & Wei, 2012, p. 58). With the application of matrices, unique words are extracted and associated with its importance through frequency count. The latent semantic space created gives essay its meaning, depending on the co-occurrence of words in the corpus used (Lemaire & Dessus, 2001). Therefore, it can only be reliable if the corpus is reliable in the first place. The weakness of this technique is that it cannot represent the actual knowledge of the students because word order, syntax, logic and other information are ignored (Landauer, Ladam& Folts, 2001).

E-rater that uses Natural Language Processing (NLP) is regarded as a revolutionary grading tool because it is based on a corpus of learner actual language. The E-rater features include "a syntactic module, a discourse module, and a topical analysis module" (Dikli, 2006, p. 54). Similar to IEA that uses information retrieval technology, E-rater applies Vector Space Model (VSM) to determine the relevance of text content (Burstein, 2003b). E-rater assumes that a good essay is resembled by other good essays and vice versa in terms of language used and content presented (Dikli, 2006).

The validity of this grading system depends on the validity of the sample grading of the corpus (Dikli, 2006).

Probably the most widely used, Intellimetric model is the very first essay scoring tools that applied Artificial Intelligent (AI) (Burstein, 2003a). It integrates AI, NLP and statistical technologies which internalises the pooled wisdom of human expert rater (Elliot, 2003). The features considered in this tool include mechanics, sentence structure, focus and unity, organisation, development and elaboration (Elliot, 2003; Dikli, 2006). Using a parsed corpus, IntelliMetric is capable of emulating the way the human brain acquires, accesses, and uses information, hence, learning the way to examine sample pre-scored essays. This system applied a non-linear and multidimensional approach to analyse essays (Elliot, 2003).

With the on-going development and enhancement of AES, the reliability of an AES system has been shown to be comparable to human marker even in high stakes examinations (Shermis, 2014).

The Novelty of AESF

Due to the lack of AES that specifically caters for Malaysian users, AESF is considered a viable, valid and reliable tool in essay marking for the Malaysian University English Test (MUET) because it is trained based on a corpus compiled using actual MUET graded essays collected from schools (Gebril & Plakans, 2014). These graded essays are scored based on the actual MUET marking criteria by experienced teachers in schools. Therefore, with valid and reliable training pool, essays graded by AESF should be more reliable than non-local commercially available AES.

Being trained using actual L2 learner corpus, AESF is also culturally sensitive as essays written by L2 will have vocabulary, structure and setting that are only familiar and acceptable by their culture, termed as 'localisation of English'. These localised English is easily intelligible by another Malaysian who is accustomed to the culture of the context (Hashim & Leitner, 2011). Endornomativity is unavoidable as English used in Malaysia is widely blended with various other languages used. For instance, borrowed words from the national language or other mother tongue are often used with or without inverted commas to make essays more vivid and realistic (Hashim & Leitner, 2011). Hence, commercially available AES will not be

able to treat such essays fairly as how an actual Malaysian marker would (Lewis, 2013).

AESF is considered state-of-the-art because it allows teachers or test administrators to have full autonomy to train, set and keep track of their students' progress. No AES can score essays topic untrained by the provider (Shermis, 2014). AESF allows the teachers to expand the marking topic by training their own topic even though this may take some time because teachers need to build up the training corpus. Teachers can upload graded essays as training set and without any additional procedure on the teacher's part. He or she can set the new topic for scoring new essays input by students. However, the number of graded essays used for training need to reach at least 200 essays before the tool can be scored reliably. This feature allows teachers to have a tool that they can use continuously with new topic added as they wish. For most AES, teachers are restricted to only pre-listed topic available for them. If the topics are not suitable for students or Malaysian context, then the AES cannot be fully utilised. Therefore, the ability to train new essays in AESF makes it a more flexible platform to utilise ICT to ease teachers' essay marking burden.

On the other hand, students will also have the autonomy to decide when they require feedback on their writing. Unlike usual word processor such as Microsoft Word that flags errors as we type, errors will only be flagged by the AESF when students request for feedback. This is similar to writing on paper where errors are not flagged immediately and students' floor of thought will not be distracted by the flagging of errors. When students request for error feedback, it means that they have written what is in their thought and is ready for feedback. Then, with the feedback flagged by AESF, students can rectify or improve on their weaknesses before continuing with their writing (Ene & Upton, 2014; Li et al., 2015). This can be done repeatedly until the students are satisfied with their performance (Attali, 2004). Some may argue that, with normal word processing, the auto correction can also be "off" but that requires additional knowledge on setting the programme and involves more indirect steps that may burden non-expert ICT users.

In addition, students are also given the opportunity to decide if they prefer to have a final score or paragraph by paragraph scoring. Final score means students will have to complete the whole essay before they submit the

essay for scoring. A holistic score will be provided to reflect the quality of the essay as a whole. In contrast, paragraph by paragraph scoring provides scores for each paragraph indicating quality of each paragraph anytime as students wish. This mimic the classroom support provided by teacher where students may ask for feedback from teachers to make sure that they are on the right track so that they can proceed writing with more confidence (Likkel, 2012).

Being networked allows higher flexibility to teachers and students in using AESF. They are not restricted by brick and mortar because AESF are not installed on computers or laptops in laboratory. AESF can be accessed anywhere via Internet connection. This overcomes the problem of insufficient computers and limited time in school to utilise ICT in education. Students can access AESF anywhere and anytime as they wish to complete their assignments. Similarly, teachers can keep track of students' progress flexibly at their convenience.

The Features and User Interface of AESF

AESF is networked, a valid link with some authentication are needed before one can get access to the system. Figure 1 shows the login page of AESF. It is designed in two modules; the teacher's and the students' module. The provision of module is set based on email registered.

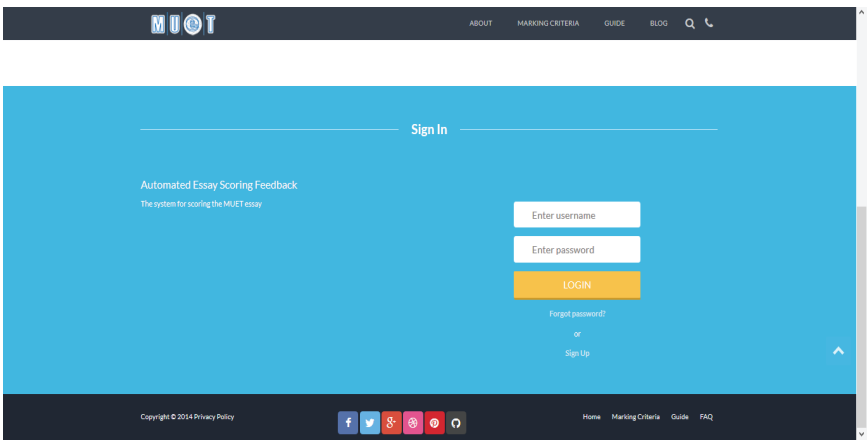

Figure 1: The Login Page of AESF

For the teacher module, the teacher can choose to collect "*gold standard*" (essays used for training AESF) or making corrections or amendments to the gold standard. Once gold standard is in placed or trained, the teacher can set the task according to the need of his/her lesson. He or she can choose the respective topics and set a timeframe for the writing assignment. Once the due date is up, students cannot submit essays or make further corrections. The user interface of the teacher module is shown in Figure 2.

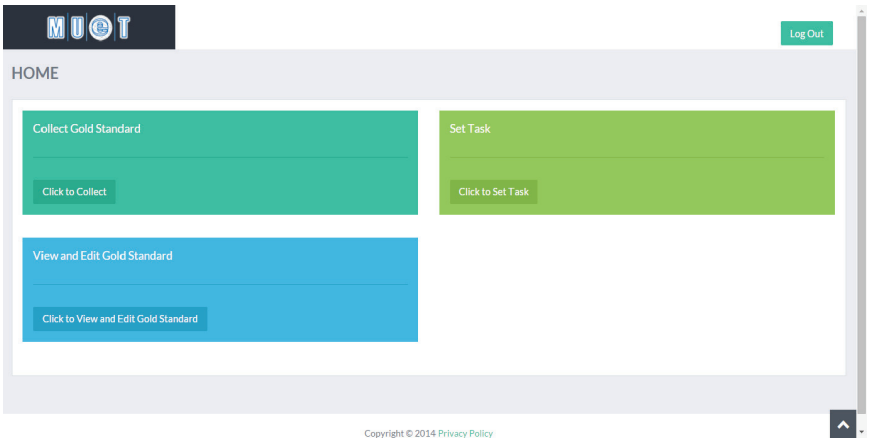

Figure 2: The Teacher Module

As for the students' module, they are shown the rubric of the essay and also the time limit set by their teachers. In this module as shown in Figure 3, students can choose to write their essay in the "full essay" option or the "paragraph by paragraph" option.

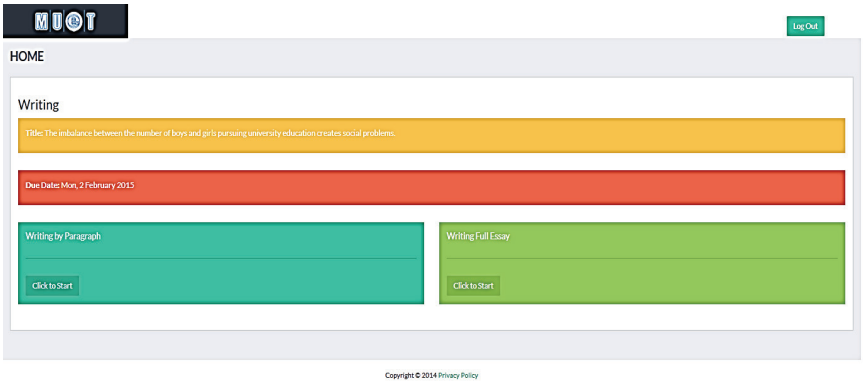

Figure 3: The Student Module

For the paragraph option, students write essays as usual by separating each paragraph using the "enter" button. Whenever students need feedback, they can hit the "preview" button. By hitting the "preview" button, AESF will automatically segment the essay into paragraph and assign a score to each paragraph together with some feedback. Students can then revise and continue writing over and over again until they are satisfied with their score before they submit their essays to their teachers.

For the full essay module, the process is more straightforward. Students will need to write the complete essay and then hit the "preview" button like the previous option. AESF will score and provide a holistic score. In addition, it also provides some general comment and some flagging of errors on the essay itself. Similarly, students can edit and re-score their essays as many times as they need before submitting their final essays to their teacher.

AESF employs the state-of-art advancement in NLP and ML to train and score essays. The AESF essentially constitutes two computing components; Essay Processor (EP) and Essay Grading Model (EGM). The EP technically is an essay analysis engine which is able to detect 10 essays properties:

1. Total word count in an essay
2. Total sentence count
3. Average words per sentence
4. Average words per paragraph

5. Average sentence per paragraph
6. Spelling error count
7. Spelling error rate (Spelling error count/Total word count in an essay)
8. Word type
9. Lexical richness (Word type/Total word count in an essay)
10. Use of noun, adjective, and adverb

The EP is built upon NLP research findings and is rather acute to extract the intended features.

Once all the sample essays are analysed and the ten features are extracted, these information will be fed into the EGM to grade student's essay. EGM essentially is built upon a machine learning algorithm, Vector Space Model (SVM) which has the capability to learn from the data given. The algorithm learns to construct a mathematical model from the input and using that to make prediction and decision of essay grade.

Referring to Figure 4, in order to grade an essay, the essay is fed into the EP, which is represented by a series of features which in turn is projected into the EGM to estimate the essay score and band.

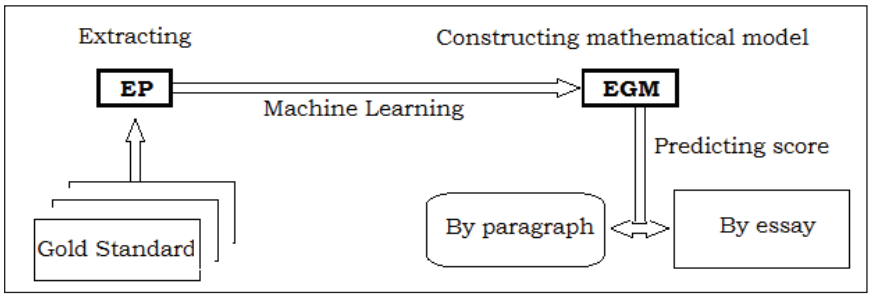

Figure 4: The Working Model of AESF

Performance Evaluation of AESF

Based on the marking rubric of MUET, essays are judged based on content, language and organisation, where content takes up 50% and the other 50% was equally shared by language and organisation. The sum of both areas suggests the final holistic score.

In order to evaluate the performance of AESF, content measurement as suggested by the lexical richness and content coherency and language proficiency as measured by the interweaving of syntactic correctness and a variety of sentence structures are combined to predict a final score. The reliability of the score prediction by AESF is compared to independent human score.

A preliminary evaluation was carried out involving an essay topic of 250 real essays, composed by different students. Each essay was graded by five participating teachers, from the distribution of Bands 1, 2, 3, and 4. The 250 essays were then fed into AESF to obtain their correspondent bands. The band from AESF is then compared to the bands given by the teachers. If the band from AESF is in agreement with the teachers, it is a hit, the other a miss. Table 1 summarizes the accuracy (number of hit/total essay) of the model to grade different band of the essay with leave-one out approach. Leave-one-out approach is a collective estimate performance of an essay predictive model trained on n−1 essay, where at each iteration, the essay being left out would be used to evaluate the model.

Table 1: Percentage of Accuracy in Scoring with AESF

Band	1	2	3	4	5	6
Accuracy	65.0%	25.7%	88.8%	0%	N/A*	N/A*

*Data is not available at the time of collection.

As we notice, the highest accuracy is on scoring Band 3 essays. This is due to the fact that the Band 3 essays are the norm in the essay collection. This is followed by Band 1 prediction, where the essays in this group demonstrate certain obvious properties: low word count, use repeated words, low number of sentences, just to name some.

However, the current system has its bottleneck at predicting Band 2 essays accurately as they demonstrate very thin borderline with Band 1 essays where most of them were miss-categorized into Band 1. In addition, Band 4 did not work as it has extremely low number of essays. In this initial study, evaluations on Bands 5 and 6 essays were not carried out due to a lack of sample essays in these two categories.

Blending AESF into the Writing Classroom

The ability of AESF in marking essays can be blended into the teaching of essay writing in common classroom as a writing tool for homework and enrichment purposes. In order not to distract and disrupt the smooth flowing of a common writing lesson, AESF will be used at the "while" writing and "post" writing stage. After the teacher has discussed the rubric and the outline of the essay based on classroom contribution, s/he will then get students to draft out their outline on paper. Either at school or at home, students may be given 1-3 days to complete and submit their essays via AESF. They are given the flexibility to write in either paragraph by paragraph or full essay option.

Once, the deadline is up, the teacher can go through the submitted essays and add on or rectify the feedback and the score assigned to each essay. From this process, the teacher will be able to identify the general mistakes that students make and identify individual students for remedial purposes apart from extracting model essays if there is any for other students to refer to. Each student's progress is also recorded each time students preview their work. A copy is saved so that the teacher will have a complete record of the students' progress and the areas that the students have come to realise and hopefully learn for future essays.

CONCLUSION

The reliability of AESF scoring can be greatly improved with the increase of the corpus size that has an equal distribution of grades. At the prototype stage, AESF demonstrated the accuracy level of 88.8% in predicting Band 3 score, hence, is confident that the same or even higher accuracy level can be achieved with bigger training sample. With wider application, more essays will feed into the system and the corpus can grow when the graded essays by the system is being moderated by human score and being added to the training database.

The value of a home grown AESF will far exceed any commercially available AESF when validity is concerned. A system that is tailor-made based on the construct of the test administered and trained using samples

of scored learners corpus of the same level ensures a valid ground for assessment. Despite the validity and reliability of AESF, it is more beneficial to fit it into the real life classroom rather than for the large scale testing of MUET simply because Malaysia does not have enough resources to administer the examination in full scale with a computer.

With AESF in the classroom, students will have a platform for self-edit and on-going practice in writing, making them more aware of mistakes and language proficiency as most L2 learners need most. The immediacy in the scoring and feedback provides more impact to students to be precise in their writing and present the best to their teachers. Teachers on the other hand can focus more on the content and development of their students' essays rather than having to correct the students' surface mistakes.

It is worth mentioning that the role of the teacher in the classroom remains important as facilitator and instructor whenever students need help in understanding the responses of AESF. No matter what, a machine remains a machine that is only to ease human activity, but not taking over the human's role.

REFERENCES

Ab Aziz, M. J., Ahmad, F. D., Abdul Ghani, A. A., & Mahmod, R. (2009). Automated Marking System for Short Answer examination (AMS-SAE) (pp. 47–51). *IEEE*. http://doi.org/10.1109/ISIEA.2009.5356500

Attali, Y. (2004). Exploring The Feedback and Revision Features of The Criterion Service. In *National Council on Measurement in Education Annual Meeting*.

Attali, Y., Burstein, J., Russell, M., & Hoffmann, D. T. (2006). Automated Essay Scoring With E-Rater V.2, The. *Journal of Technology, Learning, and Assessment*.

Burstein, Chodorow, M., & Leacock, C. (2004, Fall). Automated Essay Evaluation: The Criterion Online Writing Service. *AI Magazine*, *25*(3), 27+.

Burstein, J. C. (2003a). *Automated Essay Scoring: A Cross-disciplinary Perspective*. Routledge.

Burstein, J. C. (2003b). The E-rater Scoring Engine: Automated Essay Scoring with Natural Language Processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated Essay Scoring: A Cross-disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates.

Dikli, S. (2006). Automated Essay Scoring.*The Turkish Online Journal of Distance Education*, *7*, 49–62.

Elliot, S. (2003). Intellimetric: From Here To Validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated Essay Scoring: A Cross-disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates.

Ene, E., & Upton, T. A. (2014). Learner Uptake Of Teacher Electronic Feedback In ESL Composition. *System*, *46*, 80–95. http://doi.org/10.1016/j.system.2014.07.011.

Gebril, A., & Plakans, L. (2014). Assembling Validity Evidence For Assessing Academic Writing: Rater Reactions To Integrated Tasks. *Assessing Writing*, *21*, 56–73. http://doi.org/10.1016/j.asw.2014.03.002.

Hashim, A., & Leitner, G. (2011). Contact Expressions In Contemporary Malaysian English. *World Englishes*, *30*(4), 551–568. http://doi.org/10.1111/j.1467-971X.2011.01729.x.

IntelliMetric® Vantage Learning. (n.d.). Retrieved from http://www.vantagelearning.com/products/intellimetric/.

Jiang, J., & Wei, W. (2012). Automated Scoring Research over 40 Years: Looking Back and Ahead. *Journal of Artificial Intelligence*, *5*(1), 56–63. http://doi.org/10.3923/jai.2012.56.63.

Landauer, T.K., Ladam, D & Folts, P.W. (2001) The intelligent essay assessor: Putting knowledge to the test. *Association of Test Publishers Conference on Computer-based Testing: Emerging Technologies and Opportunity for Diverse Application*. Tuscon: Association of Test Publishers.

Lemaire, B., & Dessus, P. (2001). A System to Assess The Semantic Content Of Student Essays. *Journal of Educational Computing Research*, *24*, 305–320.

Lewis, J. K. (2013). Ethical Implementation of an Automated Essay Scoring (AES) System: A Case Study of Student and Instructor Use, Satisfaction, and Perceptions of AES in a Business Law Course. *Faculty and Staff at Digital Commons @ Salve Regina*. Retrieved from http://digitalcommons.salve.edu/fac_staff_pub/47.

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking The Role of Automated Writing Evaluation (AWE) Feedback In ESL Writing Instruction.*Journal of Second Language Writing*, *27*, 1–18. http://doi.org/10.1016/j.jslw.2014.10.004.

Likkel, L. (2012). Calibrated Peer Review Essays Increase Student Confidence in Assessing Their Own Writing. *Journal of College Science Teaching*, *41*(3), 42–47.

Mayes, R. (2014). Putting Machine Testing to the Test. *Futurist*, *48*(1), 6–8.

Page, E. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated Essay Scoring: A Cross-disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates.

Rudner, L., & Gagne, P. (2001). An Overview of Three Approaches To Scoring Written Essays By Computer. *Practical Assessment, Research & Evaluation*, *7*(26). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=26.

Shermis, M. D. (2014). State-Of-The-Art Automated Essay Scoring: Competition, Results, And Future Directions from A United States Demonstration. *Assessing Writing*, *20*, 53–76. http://doi.org/10.1016/j.asw.2013.04.001.