REVIEW OF *CRITERION*®

Title	Criterion® v. 9.1
Platforms	Windows and Macintosh
Publisher	Educational Testing Service (https://criterion.ets.org/)
Minimum software requirements	Microsoft Windows 2000, XP or Vista, Macintosh OS 9.2 or newer, Linspire or Novell Linux Desktop 9, Microsoft Internet Explorer 6.0 through 7.0, Firefox 2.0 through 3.0 or Safari, Javascript enabled, Connection speed at least 56K or higher
Support Offered	The help function provides a step-by-step introduction to the important topics on Criterion. Through a search engine, users can quickly access particular information. Additional resources for instructors are also available in the Criterion Website, including a user manual, refresher videos, and separate guidelines for elementary, secondary, higher, and international educators.
Target Language	English
Target Audience	Writing teachers, students, and language program administrators
Price	Only institutional purchase orders are accepted. Price varies depending on the number of students per session and per year (e.g., for 50-150 students, $16 per student per session and $19.10 per year), or the number of essay submissions ($5.90 for 150 essays, $3.80 for more than 2,000 essays). In addition, there are six training workshops, each of which costs $300. All new Criterion users are required to purchase at least one training session. More information available from criterionglobal@est.org

Review by Hyojung Lim and Jimin Kahng, Michigan State University

INTRODUCTION

Criterion is a Web-based learning tool that aims to support writing instruction across many age groups and several genres. The feedback provided by the software can be oriented toward English language learners and has therefore been recently marketed more specifically as an English language learning tool. Once students submit their essays to it, Criterion instantly provides holistic scores and annotated diagnostic feedback by automatically evaluating the quality of essays against, aptly, a set of criteria. It also allows follow-up writing activities, such as online peer editing, one-to-one writing conferences, and the submission of multiple revisions. The developers suggest that the primary value of this software is its efficient scoring procedure and interactive nature, thereby lightening the writing teacher's workload and amplifying opportunities for practice.

Criterion uses the e-rater® scoring engine (Burstein et al., 1998), the automated essay scoring system developed by the Educational Testing Service (ETS). The use of Criterion as an evaluator of writing has been controversial; while there is consensus on its high correlation with human raters, the high rate of its erroneous error detection has been frequently reported by researchers (Han, Chodorow, & Leacock, 2006; Tetreault & Chodorow, 2008). Despite the ongoing debate on the validity of the Criterion program, ETS (Heilman & Tetreault, 2012) reported that the e-rater rated 5.8 million scripts in 2010, and is currently utilized for more than 20 applications (e.g., Criterion, GRE, TOEFL, TOEFL Practice Online).

What is E-rater?

E-rater is an automated essay-scoring system that Criterion uses to evaluate submitted essays. The system is based on natural language processing to extract features from essays and to predict, statistically, what human raters would assign as holistic scores (Enright & Quinlan, 2010). Developed in the mid-1990s, the system has evolved to better model human scoring. Given that human raters score essays based on a given rubric, e-rater is programmed with more than 50 language features that capture aspects of the rubric. A large number of evaluating features are organized into 11 higher-level features: grammar, usage, mechanics, style, organization, development, positive features, lexical complexity (average word length, sophistication of word choice), and topic-specific vocabulary usage (scores assigned to essays with similar vocabulary, similarity to essays receiving highest scores) (Heilman & Tetreault, 2012; see Enright & Quinlan, 2010 for the features and microfeatures). E-rater analyzes the rate of errors in grammar, usage, mechanics, and style; the number and length of discourse units (e.g., thesis statement, main idea, or supporting idea); the lexical complexity (e.g., type-token ratio) and sophistication of word choice based on word frequencies; the relationship of vocabulary used to that used in top-scoring essays on the same prompt; and the length of an essay (Warschauer & Ware, 2006). E-rater scores essays by first measuring a number of features in an essay and aggregating them into feature scores. Then, the final e-rater score is generated by weighting feature scores using a regression scoring model. The model is built by processing essays scored by human raters through e-rater and using linear regression analysis to determine the optimal weighting scheme that best predicts the human ratings (Enright & Quinlan, 2010).

How Criterion Works

When creating classes initially, teachers can set several user options such as the type of feedback provided, the degree of peer review possible, and the version of the Criterion writers' handbook. The writer's handbook is an online resource tool that provides feedback definitions, example sentences, and additional lectures on good writing skills. There are five versions available: Elementary, Middle school, High school/College, English language learners (ELL), and Bilinguals (e.g., Spanish/English, Chinese/English, and Japanese/English).

Teachers design a writing assignment by clicking "assignment options" (see Figure 1) and selecting from a category (e.g., from grade 4 to GRE level), a topic mode (e.g., persuasive, informative, explanatory, informative, narrative, issue, and argumentative), and an essay topic. More than 180 essay topics and approximately 400 writing assignments are stored in the Criterion topic library for higher education. If a predetermined topic is selected, the corresponding essay prompt will automatically appear in the "enter essay prompt" box. A teacher can also create an original essay prompt by clicking "text editor" from the essay topic drop-down box and enter in the self-generated prompt. Criterion scores submitted essays by comparing each essay to a catalogue of scripts written by various international students and scored by ETS raters in the past. If a teacher uses his or her self-generated prompts, however, ETS recommends checking the appropriateness of Criterion scores by manually rating sample essays. For assessment, a teacher can set a writing-time limit: 30 minutes is the default. For the learning and writing practice, however, teachers can opt not to impose time limit constraints. As additional functions, a teacher can decide on the number of (re)submissions, the feedback categories to be displayed (traits, as they are called in Criterion—feedback categories on grammar, usage, mechanics, style, and organization & development), and assignment deadlines.

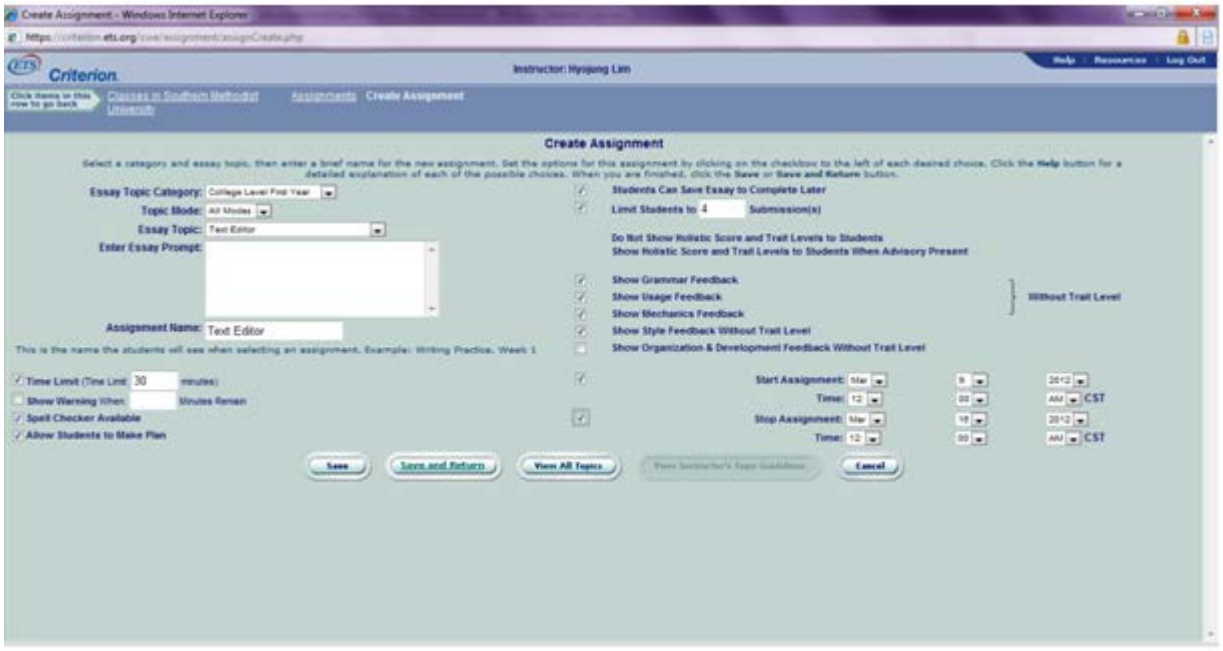

Figure 1. The "Create Assignment" page.

Teachers can view the results of the Criterion scoring and see whole-class or individual student feedback; Criterion generates summary tables and charts for the entire class, which are helpful for gaining an overall perspective of the class's performance and their collective patterns of writing errors. Criterion also lists individual students' performances, including their holistic scores and the analytic feedback regarding the five underlying categories (grammar, usage, mechanics, style, and organization & development). In addition to Criterion feedback, a teacher can make his or her own comments and provide more feedback inside the student essay by using instructor pop-up comments. Also, the teacher and student can have online (typed) dialogues to further discuss the essay and feedback, which is recorded in the comment history box next to the student's writing (see Figure 2).

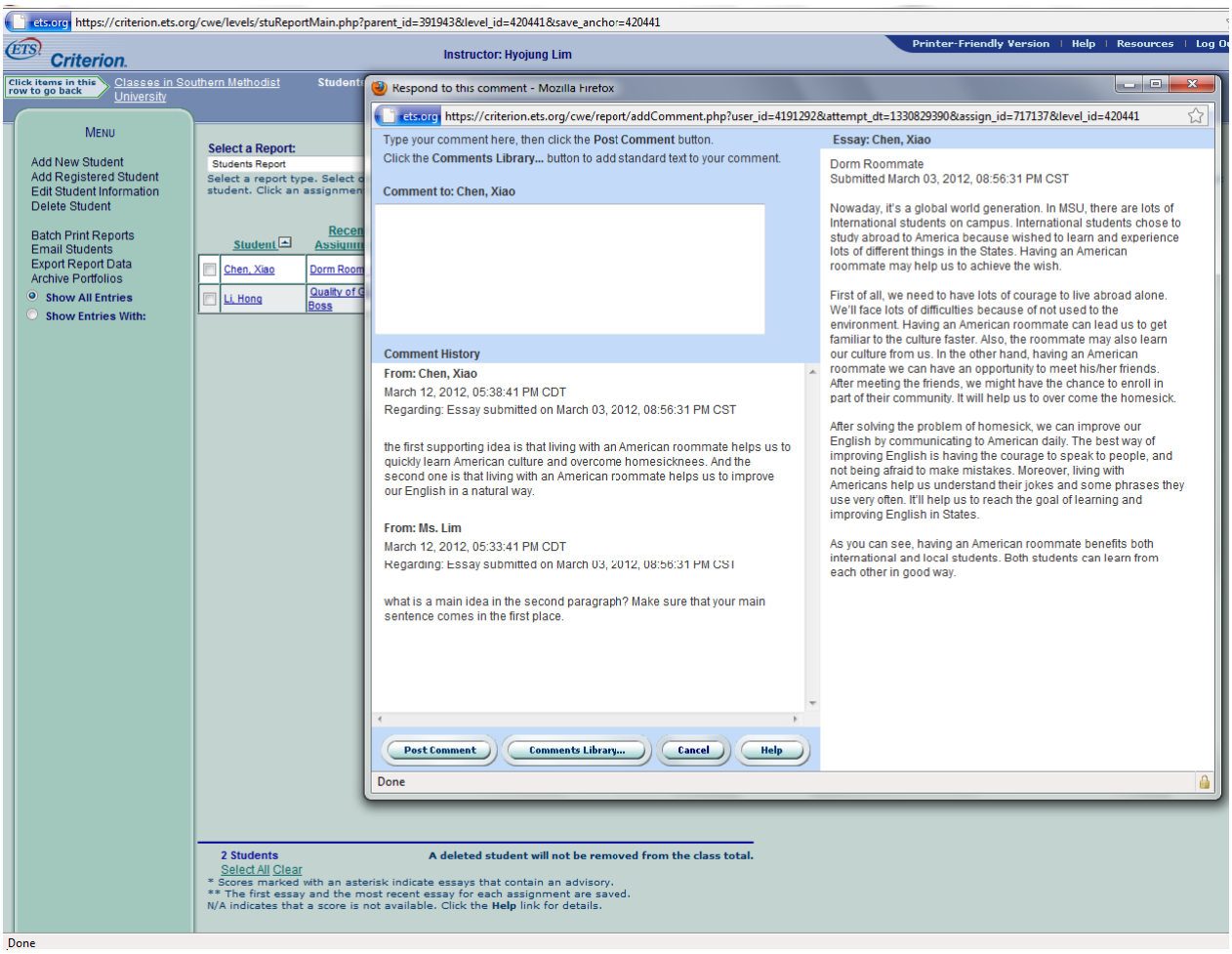

Figure 2. Teacher Comments in Criterion.

Once an instructor creates an essay assignment, students can log in and use a number of planning tools such as outline, list, idea tree, free writing, or idea web, any of which can then populate the students actual essay-writing space. Once an essay is submitted, a performance summary is generated that presents the holistic score and the number of errors and the corresponding feedback on each error (see Figures 3 and 4). The student is also provided with a score guide along with level descriptions and model essays from the catalogue of essays previously scored by ETS raters. If multiple revisions are submitted, Criterion generates a progress report on the first and the most recently submitted versions, as no more than two essays per prompt are saved in the students' portfolios.

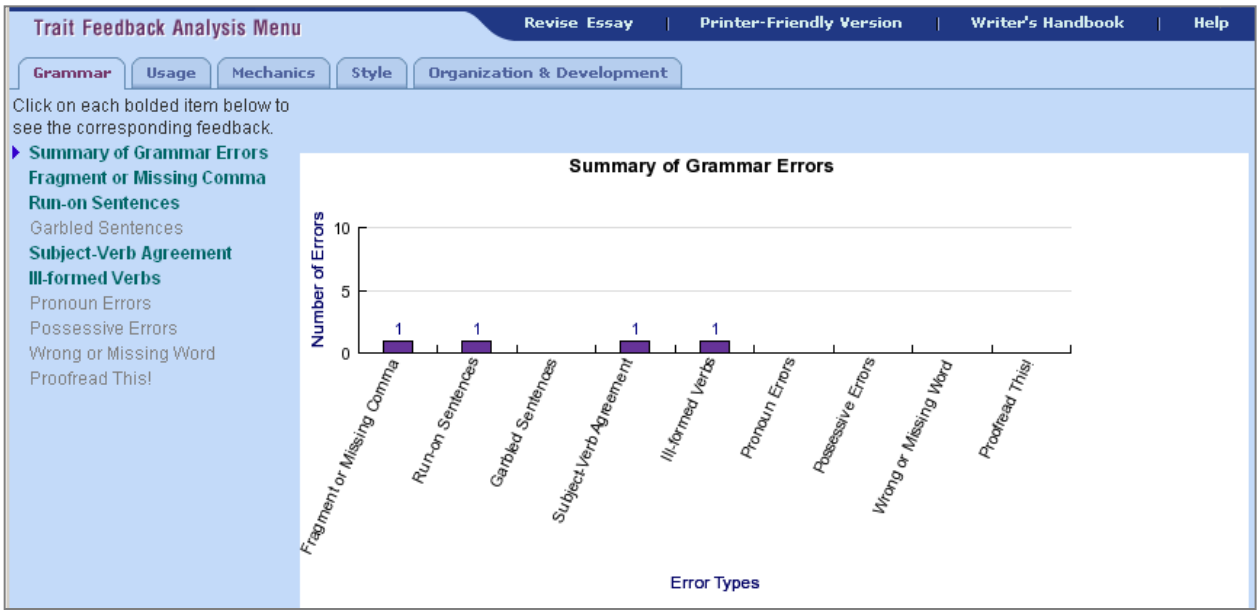

Figure 3. Trait feedback analysis on grammar.

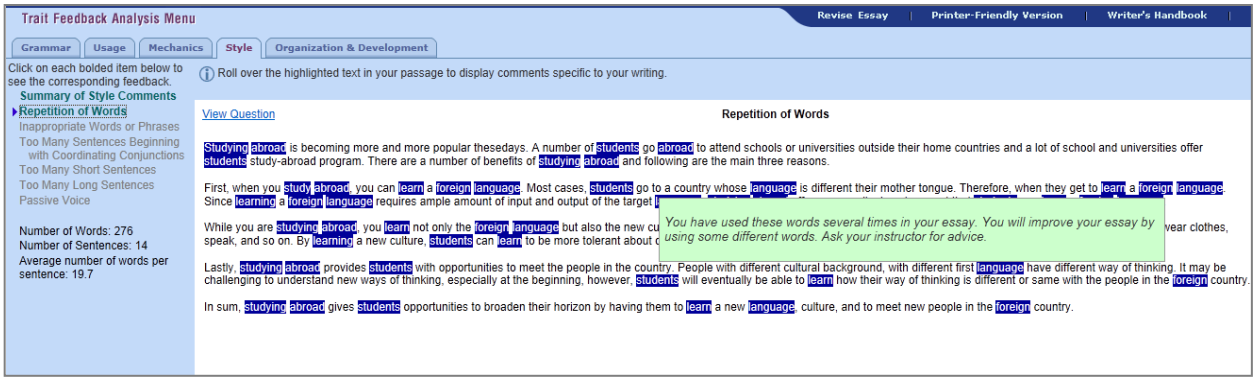

Figure 4. Trait feedback analysis on style and detailed feedback.

Criterion as a Testing Tool

In order to evaluate Criterion as a testing tool, it is crucial to understand the strengths and limitations of e-rater. E-rater has a number of strengths, including speed and automaticity. It takes only a couple of seconds to analyze a number of features in an essay and generate holistic scores. It is objective in that scoring is based on feature modules and values and is computed by statistical analysis. Another strength is the increasing amount of literature on the validation of e-rater for essay scoring. E-rater's holistic scores are claimed to be as reliable as humans' (Attali, 2007; Attali & Burstein, 2006; Enright & Quinlan, 2010). Weigle (2010) showed that the correlations between e-rater and human ratings were indistinguishable from those between two humans' ratings, suggesting that the two different rating procedures are measuring writing equally well. As for criterion-related validity, Weigle's findings demonstrated that e-rater and human scores are also comparable in regard to their relationships to non-test indicators of writing ability, such as student self-assessment and instructor assessment.

However, e-rater has several limitations. The system does not cover all the constructs of writing. It emphasizes writing quality over content, focusing on linguistic accuracy and text structures. It assesses very little in terms of argumentation or coherence. Although the system notices an essay that is written on an irrelevant topic in comparison to the given prompt, it cannot analyze argumentation, logic, or coherence as human raters do (Heilman & Tetreault, 2012). Moreover, the level of accuracy in error

detection is often not satisfactory from the perspective of teachers and learners.

Criterion as a Learning Tool

Many studies on automated scoring tools have focused mainly on system-centric evaluation—whether Criterion correctly captures errors and provides relevant feedback (Chodorow, Gamon, & Tetreault, 2010). However, to accurately evaluate the value of Criterion as a learning tool, the effect of Criterion on writers needs to be taken into account. The learning benefits of Criterion have been evidenced by a number of empirical studies. No matter if they are writing in their native language or in English-as-a second language, those who revised English papers based on Criterion feedback showed significantly fewer errors in their resubmissions, compared to no feedback groups, and thus improved their writing scores (Attali, 2004; Chodorow et al., 2010). Beneficial effects have also been reported by instructors and students; depending on the ways it is implemented, Criterion can lead to increased writing practice (Myers, 2003), can help with language use in writing (Chen & Cheng, 2008), and can draw learners' attention to linguistic features (Cheville, 2004).

Using Criterion can be beneficial for language learners, given that its interactive nature may motivate learners to be actively involved in process writing, and that ample online references embedded in the system can help learners with autonomous learning. As described earlier, students can have text-based dialogues with a teacher and with other students online; because students can read and make comments on each other's work, peer evaluation and/or editing can follow as post-writing activities. Moreover, the text-based correspondence between a teacher and a student serves as a type of one-to-one writing conference. Since the comment history box saves the online conferences, students can always refer back to them as many times as necessary to better understand the teacher's comments. Writing conferences are essential to writing development, in considering that a student can clarify his/her misunderstandings of the teacher's comments and that a teacher can provide more accurate feedback to the student (Kroll, 2001). More importantly, the online dialogues compensate for one of the major weaknesses of Criterion by allowing a teacher and student to discuss other aspects of the essay Criterion did not evaluate, such as voice, theme, logic, and content of writing. One possible flaw is that low-proficiency English learners might not be able to take much advantage of the online conference tool since they struggle with writing and responding to comments. In the future, it would be beneficial for Criterion to allow for the recording, storing, and playing of asynchronous, voiced comments.

In addition, students are given immediate access to ample referencing sources online; they can refer to a writer's handbook and model essays, if necessary. Online references may provide preemptive feedback, as opposed to diagnostic feedback that is only presented in response to student errors. Because a writer's handbook also offers additional lectures on writing skills in relation to the errors that students make, Criterion may help learners to study independently. In a similar vein, Criterion feedback provides metalinguistic explanations by naming types of errors (e.g., run-on sentences, subject-verb agreement). In addition to the commentary functions, learners can monitor their writing development through the progress report and use that information to revise further; many SLA scholars have underscored the importance of metacognitive knowledge, that is, what students know about their own cognitive processes, in terms of second language writing as well as second language development (Cotterall & Murray, 2009; Heift, 2004; Hirose & Sasaki, 1994; Schoonen, van Gelderen, Stole, Hulstijn, & de Glopper, 2011). In this regard, Criterion might prove a useful learning tool because it helps learners visualize their development and classify their writing mistakes.

While many researchers highlight what Criterion can do for writing instruction, both instructors and students should be aware of the program's limitations. They must know it does not detect all errors. As mentioned earlier, Criterion often fails to detect the errors that human raters would not miss, and its feedback on identified errors can be inaccurate. On the other hand, Criterion's incorrect feedback on errors, in the hands of strong teachers, can help learners be critical and advance their second language

learning (Swain & Lapkin, 1998). In terms of error detection, moreover, since Criterion does not deliver much information regarding the quality of learner errors (e.g., which errors are more representative of learners' language development), it is noteworthy that instructors should intervene and inform students what grammatical features learners need to work on.

Lastly, since Criterion emphasizes errors regarding surface-level linguistic features or textual forms (not voice or content), writing practice with Criterion at this time can be limited to only formulaic or mechanical writing. Criterion may discourage writers' development and use of logic and imagination (Cheville, 2004). Concerns have been voiced that the objective of Criterion appears to be accuracy and a high score, above promoting personal or communicative types of writing. For a high score, writers must conform to what ETS raters have collectively described as indicators of good writing. Thus, despite the contributions that Criterion can make to writing instruction, instructors should still act as a personal writing coach and help learners develop their individual writing styles.

CONCLUSION

Criterion is fast, automatic, and objective. It can lighten teachers' workloads and potentially amplify learner opportunities for practice. Criterion scores are highly correlated with human raters' holistic scores. For teachers and students to make the most of the program, however, they should be critical consumers; it does not evaluate content, argumentation, or coherence. Its error detection has limitations in that it misses many errors that can be detected by human raters. Despite the shortcomings, Criterion can be a useful educational tool, especially if it is used by motivated students and a well-informed writing instructor. Criterion's online references, interactive feedback, and digital records of learner performance can help augment L2 learners' metacognitive, L2-writing knowledge. But more empirical, evaluative studies of this type of software are necessary so teachers and learners can understand more concretely the best way to use Criterion.

ABOUT THE REVIEWERS

Hyojung Lim is a Ph.D. student in Second Language Studies, currently teaching an undergraduate TESOL course at Michigan State University (MSU). In the past years, she worked in the Testing Office of the English Language Center at MSU, therein assisting with English test development and administrations for ITAs and ESL/EFL students.

Email: hyojung@msu.edu

Jimin Kahng is a Ph.D. student in Second Language Studies at Michigan State University. As a teacher and a researcher she has worked on several research projects focused on second language curriculum and test development. Her primary areas of research include second language assessment, speech perception and production, and cognitive processes in second language acquisition.

Email: kahngji@msu.edu

REFERENCES

Attali, Y. (2004). Exploring the feedback and revision features of *Criterion*. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.

Attali, Y. (2007). Construct validity of e-rater in scoring TOEFL essays (ETS Research Report No. RR-07-21). Princeton, NJ: Educational Testing Services.

Attali Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *Journal of Technology, Learning, and Assessment, 4*(3) 1–30.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris M. D. (1998). Automated scoring using a hybrid feature identification technique. *Proceedings from the 36th annual meeting of the Association for Computational Linguistics*, East Stroudsburg, PA, 206–210.

Chen, C., & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94–112. Retrived from http://llt.msu.edu/vol12num2/chencheng.pdf

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *The English Journal, 93*(4), 47–52.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing, 27*(3), 419–436.

Cottereall , S., & Murray, G. (2009). Enhancing metacognitive knowledge: Structure, affordances and self. *System, 37*, 34–45.

Enright, M., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing, 27*(3) 317–334.

Han, N.-R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering, 12*(2), 115–129.

Heift, T. (2004). Corrective feedback and learner uptake in CALL. *ReCALL, 16*(2), 416–431.

Heilman, M., & Tetreault, J. (2012, March). *Using automated scoring to analyze student writing* [PowerPoint slides]. Paper presented at Georgetown University Round Table on Languages and Linguistics (GURT) 2012, Washington DC.

Hirose, K., & Sasaki, M. (1994). Explanatory variables for Japanese students' expository writing in English: An exploratory study. *Journal of Second Language Writing, 3*(3), 203–229.

Kroll, B. (2001). Considerations for teaching an ESL/EFL writing course. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed) (pp. 219–232). Boston, MA: Heinle & Heinle.

Myers, M. (2003). What can computers and AES contribute to a K–12 writing program? In M. D. Shermis & J. Burstein (Eds.). *Automated essay scoring: A cross-disciplinary perspective* (pp. 3–20). Mahwah, NJ: Lawrence Erlbaum Associates.

Schoonen, R., van Gelderen, A., Stole, R., Hulstijn, J., & de Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31–79.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal, 82*(3), 320–337.

Tetreault, J. R., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. *Proceedings of the 22nd International Conference on Computational Linguistics COLING 08, 1*(August), 865–872. Association for Computational Linguistics

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 1–24.

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.