Incorporation of Automated Writing Evaluation Software in Language Education: A Case of Evening University Students' Self-Regulated Learning in Taiwan

Bin-Bin Yu

***Abstract*—This paper reports on a preliminary investigation of university students' self-regulated learning through automated writing evaluation (AWE) software, with particular reference to evening university students in Taiwan. The purpose of the study is threefold: to examine the changes in error rates in various aspects of learner essays before and after AWE use, to compare participants' common errors with those of native speakers, and to find out possible factors that give rise to non-native speakers' common errors. This research is designed as a case study. Findings show a significant increase in the score and the length of student essays after AWE use as well as the repetition of words to be the weakest aspect of writing for both native and non-native students. Moreover, language transfer is a key factor that leads to the recurrent errors of the participants.**

***Index Terms*—Automated essay scoring, automated writing evaluation, second language writing, self-regulated learning.**

I. Introduction

With the irresistible trend of globalization and internationalization worldwide, the ability to write well in English as a lingua franca for communication in diverse fields across cultures has become an imperative in second/foreign language education. However, heavy workload of grading vast numbers of repeated drafts of student writing frequently causes some hindrance to the teaching of second/foreign language writing. In order to reduce writing instructors' workload and provide instant scores along with feedback, the software of automated writing evaluation (AWE) [1], also referred to as automated essay scoring (AES) [2], has been developed since the 1960s [3]. Nowadays, with the advancement of artificial intelligence technology, AWE systems have been developed not only for summative assessment, which is a judgment, but also for formative assessment, which is summative assessment plus feedback [4]. Students can then make use of the online AWE features to help them write and revise their essays outside the classroom for self-regulated learning [5], [6]. Thus, the instructional efficacy of AWE programs increases when their use transforms from summative assessment to formative assessment [1], [7].

The purpose of the study was to investigate learners' autonomous use of an AWE program, *Criterion* (Version 10.2), in improving English writing, with particular reference to evening, or part-time, students in Taiwan, who work during the day and attend classes in the evenings. The following three research questions guided the study:

1) Are there any changes in error rates in different aspects of student essays after AWE use?
2) Are there any differences in common errors between native speakers of English and non-native language learners?
3) What are possible factors that give rise to non-native speakers' common errors?

The next section will provide a brief review of relevant literature, which will be followed first by a description of the method used. This will then be followed by the discussion of the results and finally concluded with the implications of students' gains.

Manuscript received May 20, 2014; revised July 24, 2014.

Bin-Bin Yu is with the Department of Applied Foreign Languages, Lunghwa University of Science and Technology, Taoyuan, CO 33306 Taiwan (e-mail: bbyu@mail.lhu.edu.tw).

II. Review of Relevant Literature

A. Research on AWE

So far, much research on AWE has been centered on psychometric issues, especially on its validity, mostly by program developers [8]-[11]. The major way to validate AWE scores was to show the high correlations between computer-generated and human-rated scores. However, different results found in other studies might cast doubt on their validation [12].

Another important psychometric issue of AWE is its credibility. Some studies have revealed that the scoring systems could easily be fooled by proficient writers only if the essays were lengthy or contained certain lexico-grammatical features [13], [14]. On the other hand, reference [2] argued that although good writers could get high scores on bad essays without adequate content, it would be almost unlikely for a bad writer to get high scores on bad essays. Although the validity of AWE scoring systems remains contentious [13]-[15], the efficacy of the diagnostic feedback seems pedagogically appealing for formative learning [16].

B. Feedback and Revision

Reference [17] investigated whether the feedback report of the AWE system *Criterion* was helpful for learners, 6^{th} to 12^{th} graders in the US, in subsequent revisions of their essays. A micro-level analysis of particular error types was conducted in his study, in which writing errors identified by *Criterion* were transformed into rates by dividing them by the essay length. Effect sizes of the difference in error rates between first and last submissions were also computed. The results showed considerable variation in different error types.

Repetition of words and spelling errors were found in 93% and 78% of the essays respectively, whereas three error types were not found in the analyzed dataset, namely run-on sentences, missing articles, and preposition errors. Within thirty error types, only one, namely garbled sentences, showed a "medium" effect size (defined by [18] as 0.5) in the results, sixteen showed "small" effect sizes (defined by [18] as 0.2), and the remaining thirteen showed even smaller effect sizes. Although Reference [17] finally concluded that the learners were able to understand and attend to the error types to some significant extent, the difference between first and last submissions was, generally speaking, "small" in terms of the value of effect sizes. The current research is based on this study [17].

III. Methodology

A. Participants

This research was designed as a case study, wherein a class of 44 students enrolled in the evening session of the applied foreign languages department in a university of science and technology in Taiwan participated in the investigation. All of the participants were part-time students, who worked during the day and attended classes in the evenings. They were fourth-year English majors. Their English language proficiency was mostly at the lower-intermediate level according to a pretest. Among the participants, there were 34 or 77% female students and 10 or 23% male students. Their average age was around 24 years old, ranging from 21 to 33 years old. They were taking the required senior year writing course and using the AWE program *Criterion* during academic year 2011-2012. They had already taken three year fundamental writing courses before. Though it was their first time using *Criterion*, it was not their first time using AWE software.

B. Apparatus

Criterion is a web-based AWE program with the essay scoring engine *E-rater* developed by Educational Testing Service (ETS) in the mid-1990s. *E-rater* provides holistic scoring on a 6-point scale, with 6 being the highest score and 1 the lowest. Along with holistic scoring, *Criterion* offers diagnostic feedback in the five main aspects of writing, including grammar (e.g. fragments, run-on sentences, and subject-verb agreement), usage (e.g. wrong articles, confused words, and wrong form of words), mechanics (e.g. spelling, capitalization, and punctuation), undesirable style (e.g. repetition of words, too many short or long sentences, and passive voice), and discourse elements (e.g. introductory material, thesis statement, main ideas, supporting ideas, and conclusion) [19].

The program can allow for multiple revisions and editing. Besides, various writing resources (e.g., *Make a Plan* and *Writer's Handbook*) and editing features (e.g., *Grammar Check* and *Error Report*) have made it not only an essay assessment device but also a writing assistance tool, and even a writing learning object. Learners can then make use of the computer-generated assessment results and diagnostic advice to help them write and revise their essays as many times as they need autonomously for self-regulated learning. However, only the first and last submissions are stored in the system.

C. Procedure

The instructor implemented *Criterion* as an integrated part of her writing pedagogy. She associated the essay genres of in-class writing drills first with those of take-home writing assignments and then with those of the midterm and final examinations. The students were required to write three topics for practice with one more optional topic for extra points.

The AWE program was used for formative assessment though it also served as a writing assessment tool for the midterm and final exams with a score of 4 set as a pass threshold, which indicates that the essay achieves a "sufficient" level of communicating the writer's ideas. Besides, the instructor counted the automated scores as part of the students' actual grades, which suggests that she might have confidence that the automated scores were able to reflect students' writing performance to a reliable extent. She also allowed the students to take advantage of the automated feedback to help them reduce errors and problems in grammar, language use, and organization during their revision process even in the midterm and final examinations, which implies that she seemed to trust such feedback to be able to provide sufficient and useful information for students to improve their writing. Thus, the students should be highly motivated to write multiple drafts using the program for self-regulated learning.

D. Data Collection and Analysis

Data collected include students' writing samples in the autumn semester of the 2011-2012 academic year, along with their automated scores and feedback generated by *Criterion*. Effect sizes were computed according to [17] in order to analyze the difference between first and last submissions of an essay.

IV. Results and Discussion

The students worked on their writing assignments independently with the AWE program for formative learning. The instructor's involvement in the students' writing process was minimal, only with little consultation.

The dataset included 184 student essays. Of these, 154 essays (84%) collected were submitted more than once. This suggests that most students have made good use of the revision capabilities of the AWE system.

TABLE I: Revised and Unrevised Essays before AWE Use

	Unrevised essays (*N*=30)		First submission of revised essays (*N*=154)	
	Mean	*SD*	*Mean*	*SD*
e-rater score	4.57	0.82	3.45	1.06
Essay length (words)	326	138	220	126

30 essays (16%) were submitted only once. Table I shows that these unrevised essays on average received a higher score (i.e. 4.57) and were longer (i.e. 326 words) than the first submissions of revised essays. This might suggest that

students would not revise essays which already received a good score or passed the threshold of score 4. For the essays that were submitted more than once, only the first and last submissions were available for analysis in the system. The focus of this study was on the revised essays.

The results will be discussed in terms of the three research questions mentioned in the introductory section.

A. Research Question 1: Are There Any Changes in Error Rates in Different Aspects of Student Essays after AWE Use?

After AWE use, the students' holistic score increased from 3.45 to 4.43 on average (see Table II for more details). Table III shows the changes in the major aspects of feedback before and after AWE use. The most obvious change is essay length, which increased from 220 to 298 words. As far as grammar, usage, and mechanics are concerned, the average number of errors in the last submission was reduced in comparison with that in the first submission. All of these findings show the students' improvement in writing. However, style, rather than grammar, was the weakest aspect in student writing.

Furthermore, a micro-level analysis was conducted on the basis of [17] in order to take a closer look at the changes between the first and last submissions in the different aspects of essays. Effect size is a way of quantifying the size of the difference between two groups. Reference [18] described an effect size of 0.2 to 0.3 as a "small" effect, around 0.5 as a "medium" effect, and 0.8 to infinity as a "large" effect. Thus, the effect size of e-rater scores 0.92, as shown in Table IV, is very large, which means a large difference between the first and last submissions, in other words, the students' large improvement in score. Table IV also reveals that there was a significant increase in essay length with a medium effect size 0.62. In addition, it also shows significant decreases in grammar errors, usage errors, and style comments with medium effect sizes as well as mechanics errors with a small effect size. These also confirm the significant effects of revision on student essays.

A micro-level analysis of particular error types was further applied to individual feedback aspects, in which essay errors identified by *Criterion* are transformed into rates by dividing them by the essay length [17], as shown in Table V to Table VIII. In the tables, the first column shows the percent of essays in the first submission with errors of each type. For example, in Table V, fragments were found in 66% of the essays in the first submission. The second and third columns show the mean and standard deviation of the error rate in the first submission. The fourth column shows the mean difference in error rates between the last and first submissions. A negative difference is expected if the feedback has a positive impact. The fifth column presents the effect size of the difference in error rates, defined as the mean difference divided by the standard deviation of the error rates for the first submission. The last column presents the mean of the percent decrease of the error rates, defined as the difference in error rates divided by the error rate for the first submission. Ideally, the percent decrease would be 100%. In Table V, the students could correct 52% of fragments.

TABLE II: REVISED ESSAYS BEFORE AND AFTER AWE USE

	Mean	*SD*
Submission number	7.85	8.76
Score of first submission	3.45	1.06
Score of last submission	4.43	0.76

TABLE III: DESCRIPTIVE STATISTICS FOR MAJOR ASPECTS OF FEEDBACK

Feedback	Submission	*Mean*	*SD*
Grammar	First	4.60	4.14
(No. of errors)	Last	3.09	3.81
Usage	First	3.25	3.05
(No. of errors)	Last	2.44	3.00
Mechanics	First	4.16	7.07
(No. of errors)	Last	3.86	8.29
Style	First	23.27	17.73
(No. of comments)	Last	23.14	17.16
Essay length	First	220	126
(No. of words)	Last	298	125

TABLE IV: DESCRIPTIVE STATISTICS FOR MAJOR MEASURES

	Mean in first submission	SD in first submission	Difference between last and first sub.	SD of difference	Effect size
e-rater score	3.45	1.06	0.98*	1.12	0.92
Essay length (words)	220	126	78*	142	0.62
Grammar	0.02250	0.01821	-0.01236*	0.01907	-0.68
Usage	0.01515	0.01249	-0.00779*	0.01289	-0.62
Mechanics	0.01769	0.02360	-0.00675*	0.01727	-0.29
Style	0.12648	0.08389	-0.03797*	0.07510	-0.45

Note. Effect size is defined as difference divided by the standard deviation of first submission. *The t-test was significant at the 0.05 level, one-tailed.

TABLE V: DESCRIPTIVE STATISTICS FOR GRAMMAR ERROR RATES

Grammar	Percent of essays with errors	Mean error rate in first submission	SD of error rate in first submission	Difference in error rates	Effect size	Mean percent decrease
Fragment or Missing Comma	66%	0.00913	0.01268	-0.00475*	-0.37	52%
Run-on Sentences	54%	0.00518	0.00660	-0.00222	-0.34	43%
Garbled Sentences	7%	0.00038	0.00148	-0.00028*	-0.19	75%
Subject-Verb Agreement	32%	0.00290	0.00551	-0.00203*	-0.37	70%
Ill-formed Verbs	37%	0.00271	0.00445	-0.00183*	-0.41	68%
Pronoun Errors	3%	0.00028	0.00168	-0.00021	-0.13	77%
Possessive Errors	12%	0.00056	0.00169	-0.00020	-0.12	36%
Wrong or Missing Word	1%	0.00007	0.00058	-0.00005	-0.09	75%
Proofread This!	18%	0.00129	0.00310	-0.00078*	-0.25	61%

Note. Effect size is defined as difference divided by the standard deviation of first submission. Mean percent decrease is defined as the difference in error rates divided by the error rate for first submission. *The t-test was significant at the 0.05 level, one-tailed.

The main findings from Table V to Table VIII can be summarized as follows. The extent of the different types of errors varied considerably. Three of the error types were not found at all in the analyzed data, namely negation errors,

missing apostrophes, and inappropriate words or phrases. On the other hand, repetition of words and missing articles were found in 95% and 78% of the essays. However, these two types of errors also showed medium effect sizes (.56 and .45 respectively); that is to say, the learners could fix the errors in the subsequent versions of their essays to some significant extent. Generally speaking, there was also a significant decrease in the error rates from first to last submissions. Mature language learners seemed to be capable of solving the problems autonomously.

TABLE VI: DESCRIPTIVE STATISTICS FOR USAGE ERROR RATES

Usage	Percent of essays with errors	Mean error rate in first submission	SD of error rate in first submission	Difference in error rates	Effect size	Mean percent decrease
Wrong Article	29%	0.00187	0.00379	-0.00127*	-0.34	68%
Missing or Extra Article	78%	0.01076	0.00985	-0.00551*	-0.56	51%
Confused Words	25%	0.00164	0.00376	-0.00111*	-0.29	67%
Wrong Form of Word	2%	0.00007	0.00058	0.00001	0.02	-13%
Faulty Comparisons	1%	0.00002	0.00026	-0.00002	-0.08	100%
Preposition Error	15%	0.00068	0.00185	0.00014*	0.07	-20%
Nonstandard Word Form	2%	0.00010	0.00073	-0.00004	-0.05	37%
Negation Error	0%	0.00000	0.00000	0.00002	-	-

Note. Effect size is defined as difference divided by the standard deviation of first submission. Mean percent decrease is defined as the difference in error rates divided by the error rate for first submission. *The t-test was significant at the 0.05 level, one-tailed.

TABLE VII: DESCRIPTIVE STATISTICS FOR MECHANICS ERROR RATES

Mechanics	Percent of essays with errors	Mean error rate in first submission	SD of error rate in first submission	Difference in error rates	Effect size	Mean percent decrease
Spelling	54%	0.01423	0.02290	-0.00515	-0.22	36%
Capitalize Proper Nouns	5%	0.00048	0.00277	-0.00032	-0.12	66%
Missing Initial Capital Letter in a Sentence	12%	0.00103	0.00351	-0.00042	-0.12	41%
Missing Question Mark	3%	0.00011	0.00077	0.00003	0.03	-24%
Missing Final Punctuation	11%	0.00056	0.00179	-0.00033*	-0.18	58%
Missing Apostrophe	0%	-	-	-	-	-
Missing Comma	4%	0.00013	0.00069	-0.00005	-0.08	40%
Hyphen Error	3%	0.00019	0.00140	-0.00014	-0.10	73%
Compound Words	14%	0.00072	0.00194	-0.00028	-0.14	38%
Duplicates	5%	0.00023	0.00113	-0.00008	-0.07	37%

Note. Effect size is defined as difference divided by the standard deviation of first submission. Mean percent decrease is defined as the difference in error rates divided by the error rate for first submission. *The t-test was significant at the 0.05 level, one-tailed.

TABLE VIII: DESCRIPTIVE STATISTICS FOR STYLE ERROR RATES

Style	Percent of essays with errors	Mean error rate in first submission	SD of error rate in first submission	Difference in error rates	Effect size	Mean percent decrease
Repetition of Words	95%	0.11310	0.07718	-0.03483	-0.45	31%
Inappropriate Words or Phrases	0%	-	-	-	-	-
Sentences Beginning with Coord. Conj.	5%	0.00055	0.00260	0.00020	0.08	-36%
Too Many Short Sentences	26%	0.01127	0.02292	-0.00257	-0.11	23%
Too Many Long Sentences	15%	0.00137	0.00555	-0.00074	-0.13	54%
Passive Voice	3%	0.00019	0.00115	-0.00002	-0.02	10%

Note. Effect size is defined as difference divided by the standard deviation of first submission. Mean percent decrease is defined as the difference in error rates divided by the error rate for first submission.

TABLE IX: MAJOR MEASURES: NATIVE SPEAKERS AND NON-NATIVE SPEAKERS

	Native Speakers			Non-native Speakers		
	Mean in first submission	SD in first submission	Effect size	Mean in first submission	SD in first submission	Effect size
e-rater score	3.70	1.09	0.47	3.45	1.06	0.92
Essay length (words)	260	143	0.39	220	126	0.62
Grammar	0.0005	0.0008	-0.15	0.02250	0.01821	-0.68
Usage	0.0005	0.0009	-0.16	0.01515	0.01249	-0.62
Mechanics	0.0020	0.0027	-0.21	0.01769	0.02360	-0.29
Style	0.0186	0.0142	-0.27	0.12648	0.08389	-0.45

Note. Data for native speakers were derived from [17]. Effect size is defined as difference divided by the standard deviation of first submission.

B. Research Question 2: Are There Any Differences in Common Errors between Native Speakers of English and Non-Native Language Learners?

The comparison of native speakers' error types with those of non-native learners is presented in terms of major and individual measures. Table IX shows a comparison of major aspects of feedback. Therein, it can be seen clearly that native speakers got a higher score (i.e. 3.70), wrote a longer essay (i.e. 260 words), and made less errors in the first submission than non-native speakers. Style was found to be the weakest aspect for both native and non-native speakers. Besides,

while mechanics was the second weakness of native speakers, grammar seemed to be more difficult than the other two aspects, namely mechanics and usage, to non-native language learners. However, mature non-native language learners seemed to be more capable of revising essays than young native speakers of English.

TABLE X: COMMON ERRORS: NATIVE SPEAKERS AND NON-NATIVE SPEAKERS

Native Speakers				Non-native Speakers			
Error type	%	Effect size	Mean percent decrease	Error type	%	Effect size	Mean percent decrease
Repetition of words	93%	-0.31	22%	Repetition of words	95%	-0.45	31%
Spelling	78%	-0.22	27%	Missing articles	78%	-0.56	51%
Confused words	48%	-0.27	28%	Fragments	66%	-0.37	52%
Fragments	35%	-0.17	20%	Run-on sentences	54%	-0.34	43%
				Spelling	54%	-0.22	36%

Note. Data for native speakers were derived from [17]. Effect size is defined as difference divided by the standard deviation of first submission

Table X shows the most recurrent error types. Therein, it seems that both native and non-native speakers were in big trouble with the repetition of words. Strictly speaking, word repetition is not an error, but an essay which is full of same words cannot be counted as a good one. Missing articles and run-on sentences are unique to non-native or Chinese speakers here, which were not found in native speakers' essays. Additionally, native speakers were less skilled in spelling than non-native speakers.

C. Research Question 3: What Are Possible Factors That Give Rise to Non-Native Speakers' Common Errors?

There are at least three possible factors that could give rise to the common errors. First of all, the trouble with repetition of words might be associated with students' vocabulary size. Due to small vocabulary size, non-native language learners could not help keeping using the same words. This, however, might be accompanied by an attempt at coherence and cohesion of content in which words that appeared in the topic were kept repeating in the essay. Pronouns for "personal reference" [20] were often marked as repetition by *Criterion*.

Second, "negative transfer" [21] from Chinese construction might cause missing articles, fragments, and run-on sentences, which often happens when the differences between native language and target language are relatively great. English articles, namely 'a', 'an', and 'the', which specify the grammatical definiteness of the noun, do not exist in Chinese. In English, a fragment is not a complete sentence in which either the subject or the verb is missing. Chinese, however, is a pro-drop language in which certain pronouns as subjects or objects may be omitted when they are in a sense pragmatically inferable. Moreover, two or more complete sentences can be joined without a conjunction or punctuation in Chinese; nevertheless, this would generate an ungrammatical run-on sentence in English.

The third factor might be viewed as the other side of the same coin. That is, unfamiliarity with English construction rules could lead to even more errors.

V. CONCLUSION

This paper has reported on a preliminary investigation of evening university students' self-regulated learning through AWE use in Taiwan. The changes in various aspects of student writing before and after AWE use have been found. After AWE use, the holistic score and essay length have significantly increased. Language transfer was found to have important influence on language learning. The finding of common errors, especially those unique to the participants, namely missing articles and run-on sentences, might present an important insight into English writing pedagogy in preparing teaching materials for native speakers of Chinese. This study attended to evening university students, who are academically low achievers in Taiwan. It is of the hope that the goal of lifelong learning can be possibly achieved through their successful experience in using the AWE program for self-regulated learning.

REFERENCES

[1] M. Warschauer and P. Ware, "Automated writing evaluation: Defining the classroom research agenda," *Language Teaching Research*, vol. 10, no. 2, pp. 1-24, 2006.

[2] M. D. Shermis and J. Burstein, "Introduction," in *Automated Essay Scoring: A Cross-disciplinary Perspective*, M. D. Shermis and J. Burstein, Eds., Mahwah, NJ: Lawrence Erlbaum, 2003, pp. 13-16.

[3] E. Page, "Project essay grade: PEG," in *Automated Essay Scoring: A Cross-disciplinary Perspective*, M. D. Shermis and J. Burstein, Eds., Mahwah, NJ: Lawrence Erlbaum, 2003, pp. 43-54.

[4] M. Taras, "Assessment - summative and formative – some theoretical reflections," *British Journal of Educational Studies*, vol. 53, no. 4, pp. 466-478, Dec. 2005.

[5] D. H. Schunk and B. J. Zimmerman, *Self-Regulation of Learning and Performance: Issues and Educational Applications*, Eds., Hillsdale, NJ: Lawrence Erlbaum Associates, 1994.

[6] B. J. Zimmerman, "Self-regulated learning and academic achievement: An overview," *Educational Psychologist*, vol. 25, no, 1, pp. 3-17, 1990.

[7] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education*, vol. 5, no. 1, pp. 7-74, 1998.

[8] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2.," *Journal of Technology, Learning, and Assessment*, vol. 4, no. 3, pp. 1-30, February 2006.

[9] M. Chodorow and M. Burstein, "Beyond essay length: Evaluating e-rater®'s performance on TOEFL® Essays," Research Report No. 73, Educational Testing Service, 2004.

[10] T. Z. Keith, "Validity and automated essay scoring systems," in *Automated Essay Scoring: A Cross-disciplinary Perspective*, M. D. Shermis and J. Burstein, Eds., Mahwah, NJ: Lawrence Erlbaum, 2003, pp. 147-167.

[11] D. E. Powers, J. C. Burstein, M. Chodorow, M. E. Fowles, and K. Kukich, "Comparing the validity of automated and human scoring of essays," GRE Board Research Report No. 98-08aR, Educational Testing Service, 2000.

[12] J. Wang and M. S. Brown, "Automated essay scoring versus human scoring: A comparative study," *Journal of Technology, Learning, and Assessment*, vol. 6, no. 2, pp. 1-28, October 2007.

[13] A. Herrington and C. Moran, "What happens when machines read our students' writing?" *College English*, vol. 63, no. 4, pp. 480-499, 2001.

[14] D. E. Powers, J. C. Burstein, M. Chodorow, M. E. Fowles, and K. Kukich, "Stumping *e-rater*: Challenging the validity of automated

essay scoring," *Computers in Human Behavior*, vol. 18, no. 2, pp. 103-134, 2002.
[15] P. F. Ericsson, "The meaning of meaning: Is a paragraph more than an equation?" in *Machine Scoring of Student Essays: Truth and Consequences*, P. F. Ericsson and R. Haswell, Eds., Logan, UT: Utah State University Press, 2006, pp. 28-37.
[16] C. F. E. Chen and W. Y. E. Cheng, "Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes," *Language Learning & Technology*, vol. 12, no. 2, pp. 94-112, June 2008.
[17] Y. Attali, "Exploring the feedback and revision features of *Criterion*," presented at the National Council on Measurement in Education (NCME), San Diego, CA, April 12-16, 2004.
[18] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Hillsdale, NJ: Erlbaum, 1988.
[19] J. Burstein, M. Chodorow, and C. Leacock, "Automated essay evaluation: The *Criterion* online writing service," *AI Magazine*, vol. 25, no. 3, pp. 27-36, Fall 2004.
[20] M. A. K. Halliday and R. Hasan, *Cohesion in English*, Harlow: Longman, 1976, ch. 2.
[21] T. Odlin, *Language Transfer: Cross-linguistic Influence in Language Learning*, Cambridge: Cambridge University Press, 1989.

Bin-Bin Yu was born and raised in Taichung (literally central Taiwan), a city located in western Taiwan. She received her first degree in Russian language and first master's degree in political science both from National Chengchi University, Taipei, Taiwan, R.O.C., and obtained her second master's degree in contemporary English language and linguistics from the University of Reading, U.K. and her PhD in theoretical linguistics also from the University of Reading, U.K. in 2007.

She was an assistant professor with the Department of Applied Foreign Languages at Ta Hwa Institute of Technology. She is currently an assistant professor with the Department of Applied Foreign Languages at Lunghwa University of Science and Technology, Taoyuan, Taiwan, R.O.C. She has presented various papers at a number of important international conferences such as AERA, CALL, and TESOL, and published numerous journal papers. Her research interests mainly include discourse analysis (especially parliamentary discourse), pragmatics (in particular its interface with syntax and semantics), and second language writing (particularly learning with computer technology).

Dr. Yu is a member of TESOL (Teachers of English to Speakers of Other Languages) International Association, U.S.A. and a member of English Teachers' Association Republic of China (ETA-ROC).