Computer Assisted
Language Instruction
Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

The Short-Term and Long-Term Effects of AWE Feedback on ESL Students' Development of Grammatical Accuracy

Zhi Li[1], *Hui-Hsien Feng*[2] *and Aysel Saricaoglu*[3]

Abstract

This classroom-based study employs a mixed-methods approach to exploring both short-term and long-term effects of Criterion feedback on ESL students' development of grammatical accuracy. The results of multilevel growth modeling indicate that Criterion feedback helps students in both intermediate-high and advanced-low levels reduce errors in eight out of nine categories from first drafts to final drafts within the same papers (short-term effects). However, there is only one error reduction of statistical significance in the category of Run-on Sentence from the first drafts of the first paper to the first drafts in the subsequent papers for both levels of students (long-term effects). The findings from interviews with the participants reveal students' perceptions of Criterion feedback and help us understand the feedback effect. Implications for a more effective use of AWE tools in ESL classrooms are discussed.

KEYWORDS: AUTOMATED WRITING EVALUATION, CORRECTIVE FEEDBACK, SHORT-TERM EFFECTS, LONG-TERM EFFECTS, GRAMMATICAL ACCURACY

Corrective feedback is an important source of input for language learners as it creates teachable moments and actionable learning opportunities (Ferris, 2006; Hyland & Hyland, 2006). While responses to student writing have drawn the attention of both teachers and researchers since the 1980s, research

Affiliations

[1]Paragon Testing Enterprises, Inc..
email: zlisu2010@gmail.com

[2]Iowa State University.
email: hhfeng@iastate.edu (corresponding author)

[3]TED University, Turkey.
email: aysel.saricaoglu@tedu.edu.tr

CALICO JOURNAL VOL 34.3 2017 355–375
https://doi.org/10.1558/cj.26382

into corrective feedback in second-language (L2) writing still raises debatable questions. One such question is concerned with what role technology can play in delivering feedback and assisting English writing.

The past decade has witnessed a noticeable transformation of automated scoring tools into feedback-generating systems for pedagogical purposes. For example, e-rater from the Educational Testing Service (ETS) is incorporated in a web-based instructional automated writing evaluation (AWE) system called Criterion, and IntelliMetric from Vantage Learning supports another instructional AWE system called MY Access! A comprehensive review of AWE software can be found in Elliot et al. (2013).

AWE tools present more opportunities for feedback provision in writing classrooms, and the use of AWE tools in instructional settings is increasing rapidly (Ware & Hellmich, 2014). Meanwhile, research on AWE tools raises some concerns, such as the practices of AWE tools in classrooms (Link, Dursun, Karakaya, & Hegelheimer, 2014), the timing and quality of AWE feedback (Ferris, 2012), and, more importantly, the effectiveness of automated feedback on students' writing development (Dikli & Bleyle, 2014; Wang, 2013). Our research questions deal with the last concern.

Literature Review

Accuracy and Effectiveness of AWE Feedback

Research on AWE feedback has mainly focused on the accuracy and effectiveness of AWE feedback to help learners improve their writing. One of the common findings of such studies is that AWE feedback is not as accurate as human feedback. For example, Otoshi (2005) compared the sentence-level feedback from Criterion and two English-language instructors on 28 essays written by Japanese adults and found that human instructors detected more errors than Criterion. In a comprehensive analysis of My Access! and Criterion feedback, Chen, Chiu, and Liao (2009) reported that the accuracy of My Access! feedback was less than 10% in some error types and Criterion had an accuracy level of 70–80% in many error types. Nevertheless, both AWE tools failed to provide adequate feedback on some common errors made by learners. More recently, Dikli and Bleyle (2014) compared the feedback from Criterion with instructors' feedback on grammar, usage, and mechanics for the essays written by 14 advanced English-language learners. They concluded that instructors' feedback was superior to Criterion's in terms of quantity and quality. For example, in the category of grammar, the total number of errors identified by the instructors was 570, with an overall accuracy rate of 98.8%. By contrast, the total number of errors identified from Criterion was 94, with an overall accuracy rate of 63%.

Studies on the effectiveness of AWE feedback on students' writing development yielded some positive findings. In Wang's (2013) quasi-experimental study of Criterion in two English major classes, the students using Criterion wrote longer essays and obtained higher machine scores over the semester than those not using Criterion. In another quasi-experimental study of CorrectEnglish, an AWE system developed by Vantage Learning, Wang, Shang, and Briody (2013) found that students in the experimental group made fewer mistakes compared with their pre-tests at the end of the semester. Furthermore, students in the experimental group made significantly fewer errors of run-on sentences, fragments, capitalization, missing articles, and punctuation.

Students' Perceptions of AWE Feedback

Besides investigating the effects of AWE feedback on learners' writing development, researchers paid attention to learner perceptions. Previous studies on students' perceptions of AWE feedback have shown that student users of AWE tools tended to trust automated feedback (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Li, Link, Ma, Yang, & Hegelheimer, 2014; Rich, 2012). Some scholars attributed this trust partially to the novelty effect, i.e., learners' initial excitement with new technology (Chen & Cheng, 2008).

As for students' attitudes towards Criterion feedback, Ebyary and Windeatt (2010) used pre- and post-treatment questionnaires and found that students generally held positive attitudes after working with Criterion. With regard to the trustworthiness of AWE feedback, Dikli and Bleyle (2014) analyzed survey responses and found students trusted Criterion feedback while acknowledging its weaknesses. According to Grime and Warschauer (2010), students using MY Access! usually processed low-level feedback on grammar and word choice first and then high-level feedback on organization and development. Since low-level feedback usually promotes "superficial" revisions on the student side, students probably found revisions based on low-level feedback easier than high-level feedback.

Research questions and theoretical basis

To date, very few studies, perhaps none, have addressed the impact of AWE feedback on learners' grammatical accuracy in particular. Examining learners' grammatical errors in their drafts within one paper and across papers in one semester can shed light on the role of AWE feedback for improving learners' grammatical accuracy. Examining learners' perceptions of AWE feedback can also offer insights into the improvement in learners' grammatical accuracy. To this end, we attempted to find the answers to the following research questions:

1. Are there short-term effects of AWE feedback on ESL students' development of grammatical accuracy?
2. Are there long-term effects of AWE feedback on ESL students' development of grammatical accuracy?

3. What were the ESL students' perceptions regarding the effects of AWE feedback on their improvement of grammatical accuracy?

Methodology

Context and Participants

With approval from the Institutional Review Board (IRB) at a large US Midwestern university, a total of 135 participants were recruited from three sections of intermediate-high level (63 participants) and four sections of advanced-low level (72 participants) ESL first-year academic writing classes in the fall semester in 2011. Participants' English proficiency levels were determined with their scores on the essay writing section of the English Placement Test, an in-house test administered by the university to evaluate the need of additional ESL instruction. The majority of the participants were from China and South Korea, and their ages ranged from 18 to 20.

Three major papers were required in the intermediate-high level class: a personal essay, a process writing essay, and a reading response essay, whereas advanced-low students wrote four major papers: a personal essay, a cause-and-effect essay, a compare-and-contrast essay, and an argumentative essay. For each paper, instructors took a process-writing approach which required students to follow similar procedures: submitting first drafts in Criterion, revising drafts based on Criterion feedback, conducting peer-review and revising drafts based on peer comments, and submitting final drafts to Criterion.

Reclassification of Criterion Feedback Types

Criterion displays errors by highlighting the problematic part of a sentence and provides students with instant feedback for improvement. Students can read the feedback by clicking on each error that is highlighted. At the time of data collection in 2011, Criterion was able to detect 40 features in five categories: Grammar, Usage, Mechanics, Style, and Organization and Development. However, there are some concerns about its classification of error types (Dikli & Bleyle, 2014). For instance, it is not clear why certain types of article errors are grouped in the Grammar category while some others are grouped in the Usage category.

To address the ambiguity of error categorization in Criterion and to reduce the number of error categories for analysis, we adopted Ferris's (2006) categorization of errors, which consists of the common errors made by ESL students. Table 1 shows the mapping of Criterion feedback types to Ferris's error categories. Given our focus on students' grammatical accuracy, we limited our analyses to nine error categories based on Ferris's (2006) study: Word Choice, Verb Form, Word Form, Articles, Pronoun, Run-on Sentence, Fragment, Sentence Structure, and Subject-Verb Agreement (see the bolded error categories in Table 1).

Table 1
Comparison of Selected Criterion Feedback Type and Ferris's (2006) Error Categorization

Ferris (2006)	Criterion	**Ferris (2006)**	Criterion
Word Choice	U1_Confused Words	Spelling	G_Wrong or Missing Word
	U_Preposition Error		M_Compound Words
Verb Form	G_Ill-formed Verb		M_Fused Words
Word Form	U_Nonstandard Word Form		M_Spelling
	U_Wrong Form of Word	**Sentence Structure**	M_Duplicates
Articles	U_Missing or Extra Article		U_Negation Error
	U_Wrong Article		U_Faulty Comparisons
Pronouns	G_Pronoun Errors	Informal	S_Inappropriate Words or Phrases
	G_Possessive Errors	**Subject-Verb Agreement**	G_Subject-verb Agreement
Run-on Sentence	G_Run-on Sentence	Miscellaneous	M_Capitalize Proper Noun
Fragment	G_Fragment		G_Garbled Sentences
Punctuation	M_Hyphen Error		M_Missing Initial Capital Letter
	M_Missing Question Mark		Proofread
	M_Missing Apostrophe		S_Coordinating Conjunction
	M_Missing Comma		S_Long Sentence
	M_Missing Final Punctuation		S_Short Sentence
Verb Tense	N/A		S_Passive Voice
Singular-Plural	N/A		S_Repetititon or Words
Idiom	N/A		

Note. U = Usage in Criterion category, G = Grammar, M = Mechanics, S = Style. Only bolded error categories were analyzed in this study.

As shown in Table 1, some error types in Ferris's classification are missing in Criterion. For example, Criterion does not provide feedback on Verb Tense, Singular-Plural, and Idiom. Additionally, some re-categorized error types do not completely match the error types in Ferris's classification. For instance, Sentence Structure error in Ferris (2006) includes "missing and unnecessary words and phrases and word order problems" (p. 85), whereas the re-categorized Criterion errors for Sentence Structure only consist of Duplicates, Negation Error, and Faulty Comparison.

Data Collection

The sequential explanatory strategy was used in this mixed-methods study (Creswell, 2009), in which quantitative data (the error counts) were collected and analyzed first to reveal the short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy, and then qualitative data (semi-structured interviews) were used to examine the participants' perceptions of and experience with Criterion feedback in their writing.

Considering the fact that raw error counts across student texts are affected by the text length and thus are not comparable, we normalized the error counts using the formula suggested by Chandler (2003): (error count/essay length) × 100. This study focused on error changes in students' first drafts (D1) and final drafts (DF) in the same papers for the short-term effects, and students' first drafts in papers 1, 2, and the final paper (P1D1, P2D1, and PFD1) for the long-term effects. For the quantitative analysis, we retrieved Criterion error counts on Grammar, Usage, and Mechanics categories from these drafts. The error counts on the categories of Style and Organization & Development were not used because they are not regarded as grammar errors.

As mentioned earlier, intermediate-high ESL students wrote three papers, whereas advanced-low ESL students wrote four papers during the semester. To make comparisons across the two levels, we decided to use data from only three papers of advanced-low students, namely, Paper 1 (P1), Paper 2 (P2), and Final Paper (PF) or Paper 4.

In addition to collecting error counts from Criterion, a number of the ESL students were interviewed at the beginning and the end of the semester through convenience sampling. Out of 135 participants, 33 (18 intermediate-high and 15 advanced-low) volunteered to attend the interviews. Out of these, 20 participants joined both the first and the second interviews, while 11 participants joined only the first, and 2 only the second interview. Therefore, 53 interviews were conducted with 33 participants.

The semi-structured interviews were conducted as part of a larger research project; therefore, in this section, we only focus on the responses which are informative for our study. In the first interview, participants were asked questions regarding what types of Criterion feedback were found helpful, what feedback they did not understand, and what errors were easy for them to correct based on the feedback. In the second interview, they were asked what kind of errors they believed improved over the semester using Criterion feedback. Each interview lasted 15–20 minutes and was audio-recorded with the participants' permission. Each interviewee was given an unidentifiable research ID, such as 101c103.

Data Analysis

Considering the hierarchical nature of the data in this study, a multilevel growth model was implemented for each of the nine error types, using the mixed procedures in SPSS 21 (IBM Corp., 2012). Multilevel growth model is a type of multilevel model (also known as hierarchical linear model, mixed model, etc.) which is appropriate for nested data structures by taking sample dependence into account (Bickel, 2007). In this study, ESL students were nested within writing sections, and the students from the same section may appear more homogeneous than the students from other sections. The multilevel growth model also uses individual subjects as the grouping variable and captures individual development across time by modeling random intercepts (initial status of grammatical accuracy) and/or random slopes (developmental trajectory across assignments).

In this study, first-level units were three papers used in the ESL writing classes. They were treated as repeated measures of writing ability and were nested within ESL students. Level-1 variables included paper (P1, P2, and PF) and draft (D1 and DF). Second-level units were the ESL students enrolled or nested within each ESL writing section. No level-2 variable was involved in this study. Third-level units were the ESL writing sections, with one level-3 variable, namely writing-class level (intermediate-high level versus advanced-low level).

Following the analytical procedures recommended by Tabachnick and Fidell (2013), we started with a three-level null model or intercept-only model without any covariate or predictor (Model 1) and then proceeded with more complex models (Models 2, 3, 4, and 5) step by step to examine model fit. The first model was mainly used to calculate intraclass correlation coefficients (ICCs). A large ICC in the null model indicates a large ratio of between-group variance to within-group variance, thus confirming the necessity of using a multilevel model (Tabachnick & Fidell, 2013). As a rule of thumb, multilevel model may not be needed when an ICC is close to zero (Hayes, 2006).

Next, we added the level-1 variable, paper, as a fixed effect to the null model and compared this unconditional linear-growth model (Model 2) to the null model to determine whether the inclusion of a new variable contributed to the model fit. Then, we added another level-1 variable, draft, the level-3 variable, class level, and two interactions (draft by paper, and paper by class level) to the fixed effect, along with the level-1 variable, paper, as a random effect to test a random-intercept model (Model 3), random-slopes model (Model 4), and a full model with both random-intercept and random-slopes (Model 5). A final model was determined based on model-fit information obtained through chi-square difference tests, comparison of Akaike's Information Criterion (AIC), as well as estimated pseudo R-squared (Bickel, 2007). Short-term and long-term effects of Criterion were further examined using post-hoc pairwise comparisons, which are not reported in this paper due to space limitations.

As for the qualitative data, 53 semi-structured interviews were first transcribed verbatim in Microsoft Word. NVivo 9.0, a qualitative data analysis software developed by QSR International (Bazeley, 2007), was used to code the transcripts. The open-coding method was employed, and recurring themes were grouped into categories that were relevant to the questions asked. Representative quotes were chosen to report in our findings.

Results

In this section, we report the quantitative findings that show the effects of feedback on both intermediate-high and advanced-low ESL students' error reduction in the short term (RQ1) and long term (RQ2), and the qualitative findings that provide insights into the changes of students' errors within one paper (short-term effects) and across papers (long-term effects) throughout the semester (RQ3).

Descriptive Statistics

Overall, students at both levels had relatively more errors in the following categories: Articles, Fragment, Run-on Sentence, and Word Choice. Meanwhile, there were very low occurrences in the error categories of Word Form, Pronoun, and Sentence Structure. The descriptive statistics also revealed some differences between the two levels at the beginning of the semester (see Appendix A). For example, intermediate-high students, as expected, had more errors in six out of nine error categories than advanced-low students. As for the other three error categories in which intermediate-high level students outperformed advanced-low students (Verb Form, Word Form, and Run-on Sentence), the differences were very small.

Short-term Effects of Criterion Feedback

The results of multilevel growth models provided more information about the short-term and long-term effects of Criterion feedback on students' development of grammatical accuracy. Level-2 ICCs in the intercept-only models ranged from .070 to .348, making multilevel models advisable. However, level-3 ICCs were either fairly small in magnitude (from .007 to .017) or absent, as some of the three-level models failed to converge. Therefore, only two-level growth models were tested in the subsequent analysis. Since the same students were measured multiple times and the error terms within each student may be correlated, unstructured covariance matrix and autoregressive covariance matrix were used and compared in the modeling process. For all nine error types, the random-intercept models (Model 4) and full models with both random-intercept and random-slopes failed to converge or were recognized as statistically mis-specified (the final Hessian Matrix is not positive definite).

Therefore, the random-slopes model, with paper being a random effect, was used as a final model for each error type based on a comparison of AIC and deviance difference among subsequent models (see Appendix B). A lower AIC value indicates a better fitting model, meanwhile a significant positive difference in deviances (chi-square values) between the two models suggests a substantial improvement of model fit.

All the error types but Word Form showed significant short-term effects of Criterion, as indicated by the significant main effects of Draft. For example, the main effect of Draft for Articles was 0.628. In other words, the first drafts on average contained more article errors compared with the final drafts of the papers, controlling for other variables (see Figure 1).

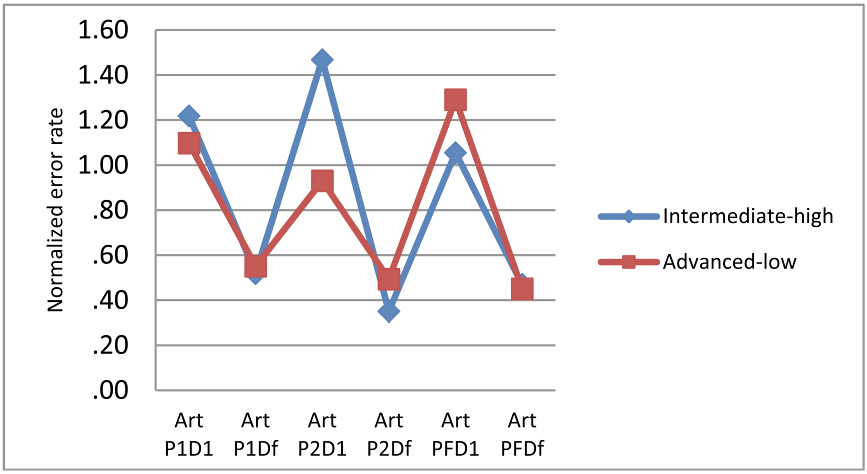

Figure 1: Changes in error rates for Articles in both levels (per 100 words).

The non-significant short-term effect for Word Form may be related with its low frequency of occurrence in P1D1, as indicated by the non-significant intercept for the error type. It is noteworthy that there was a significant interaction effect between Draft and Paper 1 for Fragment errors. Post-hoc pairwise comparisons indicated that the normalized error rate of Fragment in P1D1 was significantly higher than in P2D1 and PFDF (see Figure 2).

Long-term Effects of Criterion Feedback

As for long-term effects of Criterion feedback, significant main effects of Papers were observed in the following three error types: Fragment, Run-on Sentence, and Subject-Verb Agreement, while significant interaction effects between Paper and Class in Fragment and Subject-Verb Agreement emerged. More specifically, there was a significant decrease of error rate of Run-on Sentence

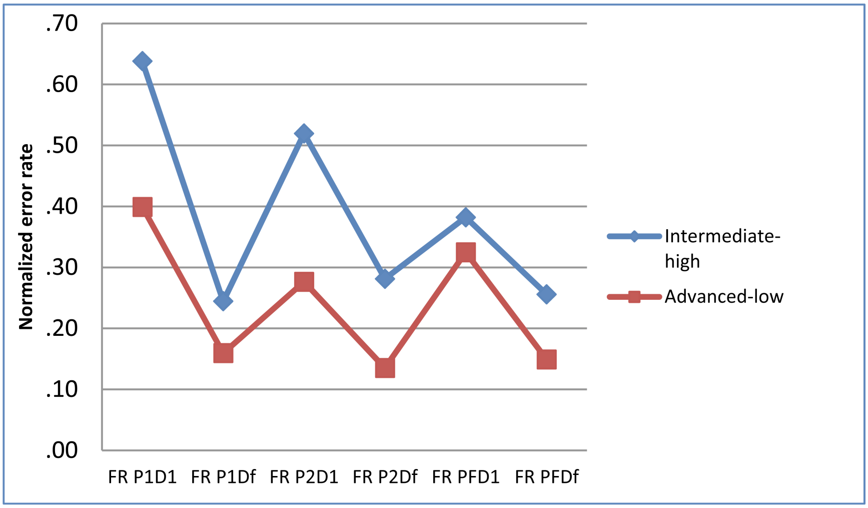

Figure 2: Changes in error rates of Fragment in both levels (per 100 words).

across the three papers (see Figure 3), which indicated a potential existence of a positive long-term effect. For Fragment, there were no significant differences among the three papers in the intermediate-high classes, while the advanced-low classes saw significantly more errors in P2 than in PF. In addition, the students in intermediate-high classes made significantly more errors than the students in advanced-low classes in P2 (see Figure 2).

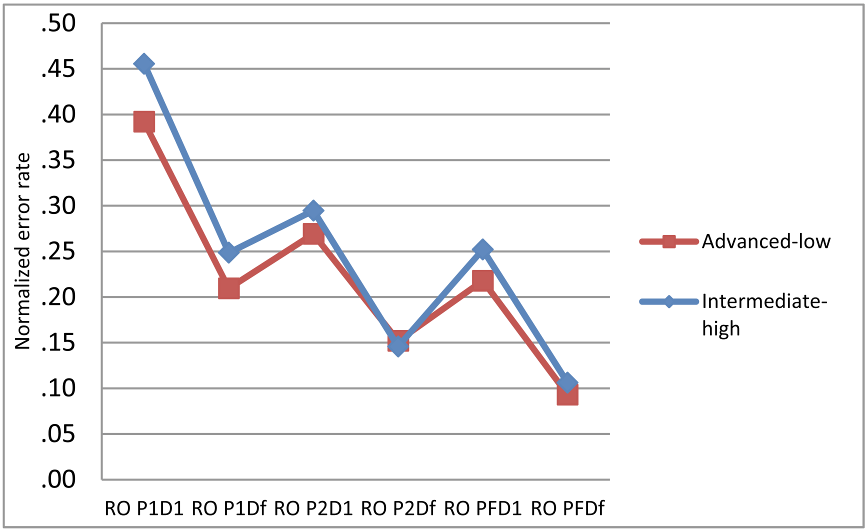

Figure 3: Changes in error rates of Run-on Sentence in both levels (per 100 words).

For Subject-Verb Agreement, the error rates decreased across the three papers in the intermediate-high classes, although the changes were not statistically significant. However, the pattern of the changes in the error rates of Subject-Verb Agreement in the advanced-low classes was different, with PF having significantly more errors than P1 and P2 (see Figure 4).

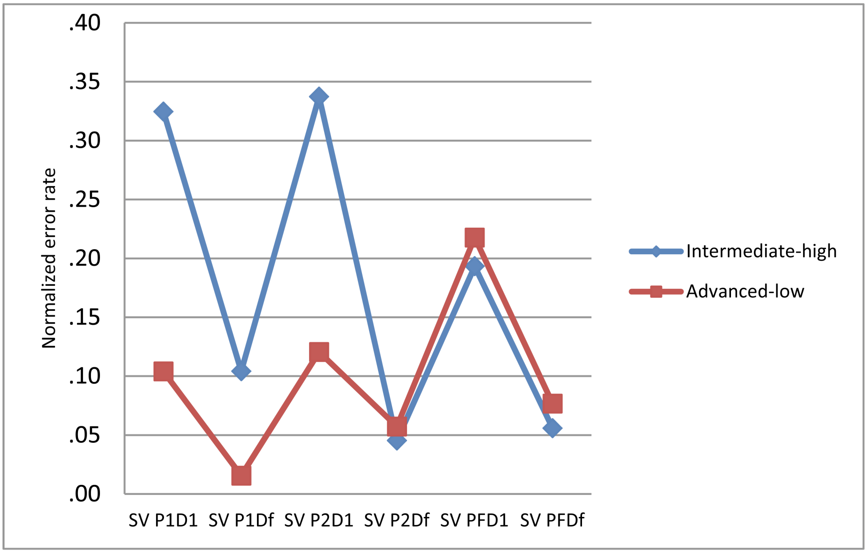

Figure 4: Changes in error rates of Subject-Verb Agreement in both levels (per 100 words).

To summarize, the quantitative analysis results revealed significant short-term effects of Criterion in eight out of nine error types, while Word Form, which had extremely low frequency of occurrence, was the only error type with non-significant short-term effect. On the other hand, only the error type of Run-on Sentence exhibited positive long-term effects of Criterion feedback in both levels of writing classes. The error types of Fragment and Subject-Verb Agreement showed somewhat different patterns of changes in the error rates across the two levels of writing classes.

ESL students' perceptions of Criterion feedback

To answer RQ3, we analyzed the interview transcripts. When asked their satisfactory level of Criterion, out of 31 students who participated in the first interview, 1 (3%) participant was very satisfied with Criterion feedback, 24 (77%) participants were satisfied, and 3 (10%) participants held a neutral attitude. Three (10%) participants did not say whether they were satisfied or not. Regarding what feedback aspects students were satisfied with, 22 (71%)

students reported that they were particularly satisfied with feedback on grammar. However, 3 (10%) students were not satisfied with Criterion because they wanted more detailed feedback. For example, one student mentioned "But, um, I wanted it, like, give me more details, like, [to] improve my sentences. Not just, like, general" (101b103).

Similarly, the responses to the question "What kind of grammar feedback was the most helpful to you?" revealed that the most helpful types of Criterion feedback were grammar (*n* = 24, 77%).

In responding to the interview question "Which categories of grammar feedback did you not understand?", 12 (39%) participants mentioned that feedback on Run-on Sentence was difficult to understand, followed by Possessive (*n* = 6, 19%), Preposition (*n* = 4, 13%), Fragment (*n* = 2, 6%), Articles (*n* = 2, 6%), Wrong Word Form (*n* = 1, 3%), and Wrong or Missing Word (*n* = 1, 3%).

According to the responses to the question "Was it easy to correct the highlighted errors in Criterion based on the feedback", 14 (45%) students stated that error correction with Criterion feedback was easy for them, whereas 3 (10%) students reported that it was not. For 10 (32%) students, it was easy sometimes, but difficult at other times due to the lack of clarity of Criterion feedback, as can be seen from one student's remarks: "Sometimes when the mistakes is a run on sentence it does not tell me what kind of mistake. I just know it is wrong but I do not know what is wrong" (101c315). The fact that the feedback did not provide details also made error correction difficult for the students: "No, I don't think. I think lots of mistake it use the same highlight and I don't make sure which one. I need more details" (101b206).

The last question asked students whether they corrected everything pointed out by Criterion. Nine (29%) participants reported they corrected every error pointed out by Criterion while 18 (58%) students explained the reasons for ignoring some of the Criterion feedback. There were three main factors related to Criterion feedback that did not lead to error correction: students did not understand the feedback (*n* = 3, 10%); they understood the feedback, but they did not know how to correct their errors (*n* = 3, 10%); and the error identification was wrong (*n* = 12, 39%).

In the second interview, we asked the interviewees if they believed that they improved their error identification skills over the semester using Criterion. Twenty-two (71%) participants reported positive opinions that using Criterion helped them improve their error identification skills. Specifically, Criterion was helpful to them for identifying the following errors: Articles (*n* = 18, 58%), Wrong Verb Form (*n* = 6, 19%), Run-on Sentence (*n* = 6, 19%), Subject-Verb Agreement (*n* = 4, 13%), Fragment (*n* = 4, 13%), Wrong Form of Word (*n* = 2, 6%), Pronoun (*n* = 2, 6%), Possessive (*n* = 1, 3%), and Faulty Comparison (*n* = 1, 3%).

Discussion and Conclusion

Our findings suggest that automated feedback from Criterion could help the ESL students reduce error rates in eight out of nine error categories in their subsequent revisions of the same paper. This finding is not surprising as it is by and large in line with the findings regarding corrective feedback in SLA studies (Li, 2010) as well as other findings about automated feedback (Stevenson & Phakiti, 2014; Wang, 2013; Wang et al., 2013). The students in this study seem to have succeeded in transferring this short-term benefit of automated feedback to a long-term gain, but only in the error category of Run-on Sentence, with no statistically significant reduction in the other eight error categories. In this section, we discuss the major findings from the interactionist perspectives and in light of the debate over the effectiveness of corrective feedback.

Short-term Effects

The short-term effect or editing effect of automated feedback can be accounted for using the Interactionist approach to SLA. Interactionists (Chapelle, 2003) view negotiation of meaning as a key to successful language acquisition through a cyclical process of input, feedback, and output. This is also the foundation of a computer-assisted language learning (CALL) environment for which Chapelle (2003) extends the interactions from between humans to between humans and computers. When language learners revise their errors detected by an AWE system, they "negotiate" with computers through their error corrections. Criterion feedback is presented as a hovering textbox on a highlighted erroneous structure. This enhanced input can easily catch students' attention. Most of Criterion's feedback is a combination of metalinguistic feedback and direct feedback (Ellis, 2008). This type of feedback is usually actionable and can promote human–computer interaction, which has the potential to contribute to short-term gains and to yield drafts with substantially fewer errors.

The positive short-term effects of automated feedback on students' grammatical accuracy can be further interpreted through students' perceptions. Overall, students held a positive view of Criterion feedback. Among the five general error categories, Grammar was the category that students found most helpful. This trustworthiness was conducive to error correction at least in the short term (Li, Link, & Hegelheimer, 2015). Using Criterion feedback for error correction resulted in significant reductions in eight error categories: Word Choice, Verb Form, Articles, Pronoun, Run-on Sentence, Fragment, Sentence Structure, and Subject-Verb Agreement. This immediate reduction in grammatical errors can probably explain learners' satisfaction with Criterion's feedback on grammar.

Long-term Effects

On the other hand, the long-term effect on Run-on Sentence errors and the absence of the long-term effects in the other error categories seem to echo Truscott's (1996, 2007) claim about the futility of grammar correction in SLA. Truscott (2007) acknowledged the effects of corrective feedback during revision, but described the effect of grammar correction on interlanguage development as "the problem of pseudolearning" because it only forms a "superficial and possibly transient form of knowledge" (Truscott, 1996, p. 344). Furthermore, drawing on Pienemann's Processiblity Theory and Teachability Hypothesis, Truscott (1996) maintained that corrective feedback is ineffective due to its failure to consider or match English language learners' developmental sequence in SLA or psycholinguistic readiness. Our findings only partially support Truscott's claim, and the differences in the effects of automated feedback on students of two proficiency levels may be relevant to their readiness in handling automated feedback. The long-term effects of automated feedback on Run-on Sentence can also be partially explained with students' interview responses, which indicated that the Run-on Sentence feedback was not easy to understand or was less useful for correction. This may reflect that students paid more attention to this type of error because it may require more mental effort to process. In that case, an in-depth processing could in turn help students become more aware of this error in subsequent writing assignments (Li, 2010).

The widely discussed U-shaped course of development in SLA can help us understand why students in this study failed to develop grammatical accuracy in eight error categories in the long run (Hyland & Hyland, 2006). In other words, initial exposure to corrective feedback enables students to make corrections and use forms properly. However, they might "regress" or temporarily forget the rules before they can internalize them and use the correct forms successfully.

We speculate that students' writing habits with Criterion could be a factor associated with the lack of long-term gains via using automated feedback. With an expectation of actionable automated feedback, students may have developed a reliance on Criterion and thus paid less attention to grammatical accuracy in their early drafts. Likewise, this mentality could divert students' attention away from language forms, especially when they focused more on meaning and content in their first drafts.

Limitations

One major inherent limitation in our research design should be acknowledged. The data came from the natural instructional settings of writing classes, and no control group was used. Additionally, some instructional differences may exist among the participatory writing classes, although same-level classes utilized the same syllabus and teaching materials. Without reference to a control group and controlling for confounding factors, the short-term and long-term effects of AWE feedback should be interpreted with caution. Since pedagogical and practical differences "can further affect student perceptions of the effectiveness of AWE in facilitating their learning of writing" (Chen & Cheng, 2008, p. 106), it would be fruitful to explore teachers' attitudes towards the use of Criterion, and how the effects of Criterion feedback change when the factors listed above are controlled through an experimental design.

Implications

Our findings show that, with the assistance of automated feedback, students at both proficiency levels could reduce most errors in the short term. On the other hand, it seems that automated feedback provided limited benefits for error deduction in the long run with an exception of the Run-on Sentence errors.

Based on the findings of this study, several implications can be drawn regarding the AWE feedback provision and AWE feedback use in ESL writing classes. An important finding from our study is that the reduction of error rates between students' drafts within and across papers differs somewhat depending on their proficiency levels. However, in current practice, different-level ESL students receive the same feedback from the same AWE tools. Previous research has shown that learner level has an influence on how effective corrective feedback is (e.g., Kennedy, 2010). AWE tool developers may want to consider a finer distinction in the content of feedback for students at different levels.

In light of our interview findings, there might be individual differences in using different types of feedback. Some students might benefit more from direct feedback than from less direct feedback. To achieve better learning outcomes, future AWE tools may consider adding an option for students to choose whether to receive direct or indirect feedback. It may also be beneficial if direct feedback is provided after learners' unsuccessful correction attempts.

Hyland and Hyland (2006) believe that "while feedback alone will not be responsible for improvement in language accuracy, it is likely to be one important factor" (p. 85). Our findings highlighted the utility of automated feedback as a valuable source for editing, which should be encouraged in ESL contexts. Meanwhile, more teachable moments could be better utilized with AWE tools in order to help students notice the gaps between their own interlanguage production and target language use, reflect on their own errors, and eventually promote development of grammatical accuracy in English writing.

About the Authors

Zhi Li is a language assessment specialist at Paragon Testing Enterprises, BC, Canada. He holds a PhD degree in applied linguistics and technology from Iowa State University, USA and an MA degree in applied linguistics from Hunan University, China. His research interests include language assessment, computer-assisted language learning, corpus linguistics, and systemic functional linguistics. He has presented his work at a number of professional conferences such as AAAL, LTRC, and TESOL. His research papers have been published in *System* and *Language Learning &Technology*.

Hui-Hsien Feng is a postdoctoral research associate at Iowa State University. She holds an MA in TESOL at the Ohio State University and a PhD in Applied Linguistics and Technology at Iowa State University. Her research interests include second language writing, automated writing evaluation, English for specific purposes, computational linguistics, and computer-assisted language learning. She has disseminated her research findings regularly in national and international conferences, including the conference of the American Association for Applied Linguistics (AAAL), the Computer Assisted Language Instruction Consortium (CALICO), Second Language Research Forum (SLRF) Conference, Symposium on Second Language Writing (SSLW), and International Conference on Computers in Education (ICCE).

Aysel Saricaoglu (Ph.D., Applied Linguistics and Technology, Iowa State University) is an assistant professor in English Language Education, TED University. She investigates academic writing with a focus on automated formative assessment, and corpus linguistics. Her work has appeared in journals such as *Computer-Assisted Language Learning* and *CALICO*.

References

Bazeley, P. (2007). *Qualitative data analysis with NVivo.* Thousand Oaks, California: SAGE Publications.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York: The Guilford Press.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, *12*(3), 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9

Chapelle, C. A. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication*. Amsterdam: John Benjamin Publishing Company. https://doi.org/10.1075/lllt.7

Chen, H. J., Chiu, T. L., & Liao, P. (2009). Analyzing the grammar feedback of two automated writing evaluation systems: My Access and Criterion. *English Teaching and Learning*, *33*(2), 1–43.

Chen, C. F., & Cheng, W. Y. (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.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Los Angeles: SAGE Publications.

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

Ebyary, K., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies*, *10*(2), 121–142.

Elliot, N., Gere, A. R., Gibson, G., Toth, C., Whithaus, C., & Presswood, A. (2013). Uses and limitations of automated writing evaluation software. *WPA-CompPile Research Bibliographies*, *23*. Retrieved from http://comppile.org/wpa/bibliographies/Bib23/AutoWriting Evaluation.pdf

Ellis, R. (2008). A typology of written corrective feedback types. *ELT Journal*, *63*(2), 97–107. https://doi.org/10.1093/elt/ccn023

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A Multi-Site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6). Retrieved from http://www.jtla.org

Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524742.007

Ferris, D. R. (2012). Technology and corrective feedback for L2 writers: Principles, practices, and problems. In G. Kessler, A. Oskoz, and I. Elola (Eds.), *Technology across writing contexts and tasks*. CALICO Monograph. San Marcos, TX: CALICO.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, 32, 385–410. https://doi.org/10.1111/j.1468-2958.2006.00281.x

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*, 83–101. https://doi.org/10.1017/S0261444806003399

IBM Corp. (2012). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Kennedy, S. (2010). Corrective feedback for learners of varied proficiency levels: A teacher's choices. *TESL Canada Journal, 27*(2), 31–50. https://doi.org/10.18806/tesl.v27i2.1054

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. https://doi.org/10.1016/j.jslw.2014.10.004

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60*(2), 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

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. https://doi.org/10.1016/j.system.2014.02.007

Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards better ESL practices for implementing automated writing evaluation. *CALICO Journal, 31*(3), 323–344. https://doi.org/10.11139/cj.31.3.323-344

Otoshi, J. (2005). An analysis of the use of Criterion in a writing classroom. *The JALT CALL Journal, 1*(1), 30–38.

Rich, C. S. (2012). The impact of online automated writing evaluation: A case study from Dalian. *Chinese Journal of Applied Linguistics, 35*(1), 63–79. https://doi.org/10.1515/cjal-2012-0006

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (4th ed.). Boston: Allyn and Bacon.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*, 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255–272. https://doi.org/10.1016/j.jslw.2007.06.003

Wang, P. (2013). Can automated writing evaluation programs help students improve their English writing? *International Journal of Applied Linguistics & English Literature, 2*(1), 6–12. https://doi.org/10.7575/ijalel.v.2n.1p.6

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

Ware, P., & Hellmich, E. (2014). CALL in the K–12 context: Language learning outcomes and opportunities. *CALICO Journal, 31*(2), 140–157. https://doi.org/10.11139/cj.31.2.140-157

Appendix A

Table 2

Mean and Standard Deviation of Normalized Error Rates across Papers and Class Levels

		Paper 1 (P1)		**Paper 2 (P2)**		**Final Paper (PF)**	
		First Draft	*Final Draft*	*First Draft*	*Final Draft*	*First Draft*	*Final Draft*
Word Choice	IH[a]	0.33 (0.39)[b]	0.18 (0.29)	0.26 (0.33)	0.08 (0.16)	0.24 (0.27)	0.12 (0.30)
	AL	0.29 (0.26)	0.15 (0.20)	0.28 (0.28)	0.17 (0.26)	0.35 (0.32)	0.16 (0.25)
Verb Form	IH	0.20 (0.29)	0.05 (0.16)	0.21 (0.28)	0.03 (0.09)	0.21 (0.29)	0.04 (0.09)
	AL	0.21 (0.25)	0.09 (0.17)	0.17 (0.21)	0.06 (0.13)	0.13 (0.16)	0.05 (0.11)
Word Form	IH	0.00 (0.00)	0.00 (0.03)	0.02 (0.14)	0.00 (0.00)	0.02 (0.09)	0.01 (0.05)
	AL	0.01 (0.03)	0.00 (0.02)	0.01 (0.05)	0.01 (0.05)	0.00 (0.04)	0.00 (0.02)
Articles	IH	1.11 (1.28)	0.55 (0.70)	1.28 (1.11)	0.41 (0.55)	1.15 (0.73)	0.55 (0.64)
	AL	1.04 (0.62)	0.57 (0.67)	0.96 (0.60)	0.56 (0.72)	1.20 (0.83)	0.54 (0.74)
Pronoun	IH	0.07 (0.19)	0.02 (0.07)	0.05 (0.15)	0.02 (0.08)	0.06 (0.13)	0.02 (0.09)
	AL	0.05 (0.11)	0.02 (0.06)	0.04 (0.07)	0.03 (0.08)	0.07 (0.10)	0.02 (0.06)
Run-on Sentence	IH	0.41 (0.45)	0.29 (0.39)	0.24 (0.33)	0.18 (0.35)	0.23 (0.26)	0.11 (0.17)
	AL	0.45 (0.40)	0.30 (0.42)	0.35 (0.42)	0.20 (0.33)	0.28 (0.27)	0.16 (0.24)
Fragment	IH	0.56 (0.66)	0.24 (0.18)	0.52 (0.61)	0.28 (0.24)	0.37 (0.30)	0.24 (0.23)
	AL	0.43 (0.63)	0.21 (0.29)	0.25 (0.27)	0.16 (0.25)	0.39 (0.41)	0.25 (0.42)
Sentence Structure	IH	0.03 (0.11)	0.01 (0.06)	0.01 (0.05)	0.01 (0.09)	0.03 (0.09)	0.02 (0.07)
	AL	0.01 (0.05)	0.01 (0.03)	0.01 (0.04)	0.00 (0.02)	0.03 (0.09)	0.01 (0.05)
Subject-Verb Agreement	IH	0.28 (0.35)	0.10 (0.23)	0.25 (0.38)	0.05 (0.14)	0.17 (0.26)	0.07 (0.16)
	AL	0.10 (0.16)	0.05 (0.17)	0.12 (0.22)	0.07 (0.13)	0.26 (0.39)	0.09 (0.17)

Notes. [a] IH = Intermediate-high classes ($n = 63$), AL = Advanced-low classes ($n = 72$). [b] The value in parenthesis is the standard deviation.

Appendix B

Table 3

Summary of Parameter Estimates in the Final Models (Random-Slope Models) for Nine Error Categories

Parameters	Word Choice	Verb Form	Word Form	Articles	Pronoun	Run-on Sentence	Fragment	Sentence Structure	Subject-Verb Agreement
Fixed effects									
Intercept (SE)	0.172** (0.032)	0.029 (0.021)	-0.0006 (0.006)	0.557** (0.086)	0.023* (0.011)	0.161** (0.030)	0.257** (0.042)	0.015 (0.009)	0.100** (0.027)
Draft = 1 (SE)	0.157** (0.027)	0.113** (0.022)	0.007 (0.006)	0.628** (0.080)	0.04** (0.012)	0.118** (0.027)	0.127** (0.046)	0.015* (0.006)	0.148** (0.029)
Paper = 1 (SE)	-0.028 (0.044)	0.053 (0.028)	0.005 (0.008)	-0.013 (0.126)	-0.012 (0.017)	0.157** (0.047)	-0.074 (0.062)	-0.011 (0.012)	-0.077* (0.036)
Paper = 2 (SE)	-0.021 (0.042)	0.009 (0.030)	0.004 (0.009)	-0.120 (0.115)	-0.001 (0.015)	0.067 (0.044)	-0.140* (0.057)	-0.010 (0.010)	-0.074* (0.036)
Class = IH (SE)	-0.072 (0.043)	0.039 (0.026)	0.015 (0.009)	-0.020 (0.114)	-0.002 (0.013)	-0.057 (0.039)	-0.018 (0.052)	0.003 (0.012)	-0.060 (0.034)
Draft =1 × Paper = 1 (SE)	-0.008 (0.038)	0.027 (0.032)	-0.007 (0.008)	-0.113 (0.115)	0.000 (0.017)	0.021 (0.039)	0.134* (0.067)	-0.003 (0.010)	-0.044 (0.041)
Draft = 1 × Paper = 2 (SE)	-0.006 (0.039)	0.031 (0.032)	0.004 (0.008)	-0.0004 (0.116)	-0.022 (0.017)	-0.002 (0.040)	0.034 (0.067)	-0.012 (0.010)	-0.023 (0.042)
Paper = 1 × Class = IH (SE)	0.099 (0.059)	-0.068* (0.034)	-0.017 (0.010)	0.010 (0.169)	0.018 (0.022)	-0.015 (0.065)	0.099 (0.078)	0.0132 (0.016)	0.173** (0.044)
Paper = 2 × Class = IH (SE)	0.004 (0.055)	-0.031 (0.037)	-0.013 (0.012)	0.103 (0.146)	0.006 (0.019)	-0.036 (0.058)	0.217** (0.068)	0.001 (0.013)	0.117** (0.043)

Random effects									
Residual	0.043** (0.003)	0.030** (0.002)	0.002** (0.000)	0.386** (0.030)	0.008** (0.001)	0.045** (0.004)	0.131** (0.010)	0.003** (0.000)	0.051** (0.004)
Variance of Paper 1		0.016** (0.004)	0.000 (0.000)	0.313 ** (0.067)	0.005** (0.001)	0.113** (0.017)	0.082** (0.020)	0.002** (0.001)	
Covariance of Papers 1 and 2		0.003 (0.003)	-0.657** (0.000)	0.206** (0.043)	0.000 (0.001)	0.062** (0.012)	0.022* (0.011)	0.000 (0.000)	
Variance of Paper 2		0.011** (0.004)	0.002** (0.000)	0.166* (0.051)	0.002* (0.001)	0.087** (0.014)	0.023 (0.013)	0.000 (0.000)	
Covariance of Paper 1 and Final Paper		0.009** (0.003)	-0.469** (0.000)	0.018 (0.042)	0.000 (0.001)	0.028** (0.008)	0.025* (0.011)	0.000 (0.000)	
Covariance of Paper 2 and Final Paper		0.003 (0.002)	0.001** (0.000)	0.060 (0.037)	0.001 (0.001)	0.027** (0.007)	0.02* (0.009)	0.000 (0.000)	
Variance of Final Paper		0.005 (0.003)	0.001** (0.000)	0.204** (0.054)	0.001 (0.001)	0.025** (0.006)	0.017 (0.014)	0.003** (0.001)	
AR1 rho	0.352** (0.107)								0.883** (0.199)
Level 2 ICC	.181	.105	.095	.137	.038	.348	.121	.070	.113
Level 3 ICC	.009			.007		.005	.014		.017
Model fit									
Pseudo R^2	38.5%	24.7%	33.2%	35.6%	22.8%	46.3%	18.2%	33.3%	13.0%
Deviance difference (df)	106.03** (12)	103.99** (13)	102.28** (13)	152.35** (12)	20.95 (13)	176.99** (12)	50.74** (12)	75.32** (13)	69.98** (12)
AIC	145.76	-271.56	-2168.18	1630.21	-1192.48	293.18	757.65	-1906.5	34.74

Note. Maximum likelihood estimation was used in all models. The covariance structure was set as unstructured, except for Word Choice and Subject-Verb agreement, in which autoregressive error covariance structure was used. SE = standard error, AR1 = autoregressive error covariance, ICC = Intraclass correlation, AIC = Akaike's Information Criterion