Castledown

OPEN ACCESS

Technology in Language Teaching & Learning
ISSN 2652-1687
https://www.castledown.com/journals/tltl/

Technology in Language Teaching & Learning, 6(3), 1–18 (2024)
https://doi.org/10.29140/tltl.v6n3.1536

Automated Feedback on Fluency and Complexity Measures for L2 Academic Paragraphs: Student Perspectives and Impacts on Human-rater Scores

EMILY MACFARLANE[a]

RYAN SPRING[b]

[a]*Tohoku University, Japan*
macfarlane.emily.b4@tohoku.ac.jp

[b]*Tohoku University, Japan*
spring.ryan.edward.c4@tohoku.ac.jp

Abstract

While complexity, accuracy, and fluency (CAF) measures are known to correlate with L2 writers' scores, less is known about the effectiveness of providing automated feedback based on these measures for improving writing performance. Furthermore, it remains unclear if improvements in specific CAF measures correspond to improvements in human-rater scores. Finally, the trade-off hypothesis predicts that all three components of CAF cannot be improved at once, so providing multiple CAF measures to students at the same time might cause cognitive overload and overwhelm students, reducing their ability to uptake the feedback. To examine these issues, a simple paragraph feedback tool was developed to provide input on number of words, vocabulary variety, supporting detail markers, and sentence length. The tool was implemented repeatedly in L2 academic paragraph writing lessons with 124 students. Improvement was assessed through complexity, accuracy, and fluency (CAF) measures and human-rater scores, while also gathering student opinions. Results showed significant improvements in number of words, sentence length, and supporting detail marker usage from pre- to post-treatment with improvements in number of words and vocabulary variety found to be the strongest predictors of improvements in human-rater scores. Students reported finding vocabulary variety feedback most helpful, while supporting detail marker feedback was perceived as most confusing. Notably, students reported very low cognitive load overall. This article also discusses pedagogical implications of these findings for L2 writing instruction and automated feedback implementation.

Keywords: CAF measures, automated feedback, L2 writing, cognitive overload

Introduction

Formative feedback is important for developing writing ability in general and is an important issue in teaching second language (L2) writing (e.g., Benali, 2021; Ellis et al., 2008; Li, 2021; Suzuki et al., 2019). Though several studies have shown that personalized teacher feedback can lead to improvement (Shintani et al., 2014; Suzuki et al., 2019), many universities in the Japanese English as a Foreign Language (EFL) context have large class sizes and teaching staff do not always have the time to give personalized feedback in a timely and effective manner. One way to overcome this problem is to offer automated feedback to students, which is prompt and can be somewhat personalized with little extra effort on the part of teachers (e.g., Carless, 2020; Li, 2021). However, there is still much disagreement regarding how useful automated feedback is, the most useful types of automated feedback, and how to best implement it practically in an EFL learning context. For example, many studies focus on providing feedback on grammatical accuracy (e.g., Gao & Ma, 2022; Suzuki, 2012; Tomen, 2022), but other studies suggest that automated grammar correction is less helpful than other types of automated feedback (e.g., Guo et al., 2021; Gao & Ma, 2022; Li, 2021). Furthermore, some automated systems provide a large amount of feedback (e.g., McNamara et al., 2014; Nguyen & Tran, 2023), but there are concerns that providing too much feedback to students at once may be too cognitively demanding and prevent uptake (e.g., Akin & MurrellJones, 2018; Nawal, 2018; Nguyen & Tran, 2023). Finally, though many studies show that particular complexity and fluency measures correlate highly with rater scores (e.g., Lu, 2010; 2012; Kyle, 2016; Kyle & Crossley, 2018; Spring & Johnson, 2022), and tend to increase in tandem with increases in general proficiency (e.g., Wolfe-Queintero et al., 1999), there are fewer studies that show that asking students to attempt to improve these particular measures results in them actually producing better human-rater scores. One exception are the works of Rastgou (2022; 2024) who looked at various feedback conditions in L2 writing and suggested that simultaneous and consistent feedback on a few areas, including CAF, can help students to improve comprehensively. However, his studies did not look at automated feedback on complexity or fluency measures. Therefore, this study seeks to add to the discussion by examining the use of an automated feedback system that provided Japanese EFL learners with four specific types of CAF measure-based feedback on an opinion writing task as part of an academic reading and writing course. Specifically, it examines pre- and posttest data, survey data, and the relationship between the two to evaluate the effectiveness of the automated feedback, students' perceptions towards the feedback and the cognitive demands of the feedback.

Background

Automated Feedback

Automated feedback on L2 writing assignments has become increasingly popular partly because corrective feedback is generally agreed to help L2 learners improve their writing ability (Benali, 2021; Ellis et al., 2008; Rastgou, 2024; Suzuki et al., 2019), but teachers are often too busy to give personalized feedback to each individual student (Niu et al., 2021; Shang, 2022). However, there is some concern about the appropriateness of automated feedback. For one, many of the studies focus on the automated feedback of primarily grammatical errors (e.g., Guo et al., 2021; Gao & Ma, 2022; Shi & Aryadoust, 2024; Tomen, 2022) with studies such as Rastgou (2024) finding that automated feedback therefore tends to only push learners to improve their grammatical accuracy. Furthermore, many recent studies have suggested that automated feedback is best when combined with other types of feedback such as that from teachers and peers (Nguyen & Tran, 2023; Niu et al., 2021; Shang, 2022; Zhang & Zhang, 2022) with some suggesting that isolated automated feedback is undesirable (Cen & Zheng, 2024). Finally, there is some concern regarding how much the feedback is converted into uptake (Gao & Ma, 2022; Shang, 2022).

Regardless of the concerns, many studies are also positive towards automated feedback showing that students are generally favorable towards it, especially due to its instantaneous and potentially individualized nature (Shang, 2022). Automated feedback allowing students to immediately see their scores increase, not only motivates students and increases confidence (Fan & Ma, 2022) but it has been suggested that the immediate feedback encourages students to try new phrases and structures (Yang et al., 2024). A meta-analysis of automated feedback by Lu et al. (2021) suggested that it has only a small to medium effect and that it is inferior to teacher feedback, but new works such as Li (2021) also point to the possibilities of automated feedback as technology improves and is tailored to the specific needs of the learners. Therefore, it seems that there is considerable research that suggests automated feedback can help L2 learners improve their writing, but that there are many factors that might impact its effectiveness, such as the appropriateness, level of the students, and whether it is provided in tandem with other types of feedback (e.g., Benali, 2021; Lu et al., 2021).

CAF Measures

Complexity, accuracy, and fluency (CAF) measures have long been suggested to correlate with L2 proficiency, especially with L2 production ability. Since much of the seminal work when such measures were often calculated by hand (e.g., Wolfe-Quintero et al., 1998), software has been created to produce CAF measures that have been reliably shown to correlate with L2 production performance (i.e., Lu, 2010; 2012; Kyle, 2016; Kyle & Crossley, 2018; Kyle et al., 2018). These programs have been further improved through the implementation of ever-more accurate parsers built on Large Language Models (LLMs) (e.g., Kim & Crossley, 2018; Crossley et al., 2019; Spring & Johnson, 2022), and to look for other aspects related to CAF such as cohesion (Kim & Crossley, 2018; Crossley et al., 2019), stance (Eguchi & Kyle, 2023), and discourse markers (Spring, 2023). These measures can be very useful not only for automated writing evaluation, but also potentially for automated writing feedback because they are generally not constrained by topic. For example, many automated writing evaluation models use the bag-of-words approach, which create models based on words and linguistic chunks that tend to score highly, which proves to be highly accurate, but creates a model that cannot be used on any other topic (e.g., Ding & Zou, 2024). Conversely, CAF measures deal with linguistic features (e.g., the number of words written, the average length of syntactic units, etc.) and are therefore more versatile when used across topics and data sets (i.e., many CAF measures correlate to proficiency similarly across different data sets: Spring & Johnson, 2022; Lu, 2012; Kyle, 2016; etc.).

However, while CAF measures do seem to correlate to learners with higher ability, most of the studies that validate these measures do so with data taken from a single data set of learners, i.e., a snapshot of the learner's responses and proficiency at the time of writing. Though some longitudinal studies have shown that CAF measures improve over time along with proficiency (e.g., Jiang et al., 2019; Vercellotti, 2017), and Rastgou (2022; 2024) suggests that providing various types of feedback including CAF related feedback helps improve students' writing, there are no studies that we are aware of that utilize automated CAF measures as feedback and track changes in the CAF measures as a result. Furthermore, while many CAF measures seem to correlate with L2 proficiency, it is still unclear if students who focus on boosting these measures will also improve in the eyes of human raters, or if drastically improved overall L2 proficiency is instead the driver of any improvement in CAF measures.

Cognitive Load and Feedback

Cognitive load theory is based on the idea that the human cognitive system has finite processing capacity, so when the demands placed on working memory by tasks are too high, and this capacity is exceeded, learning may be hindered (Sweller, 1998). During a task, the cognitive load is the overall demand placed on the student's cognitive system and therefore it is a multifaceted concept which

also includes mental exertion, weariness, and emotional strain (Fyffe et al. 2015). There are three main types of cognitive load: extraneous or extrinsic, intrinsic, and germane (Sweller et al, 1998). The Intrinsic load is created by the task difficulty, or complexity of the task content, in relation to the individual's expertise. Extrinsic load is caused by unnecessary cognitive processing due to poor task design in the presentation of the task. Finally, Germane load refers to the students' own concentration levels and how they analyze information (Akin & MurrellJones, 2018). Though each of these can be difficult to measure (e.g., de Jong, 2010; Kalyuga, 2011), each is considered an important part that can potentially affect learners' ability to conduct a task or integrate feedback.

When doing an L2 writing task, the writing process already requires significant cognitive effort requiring students complete focus and concentration with the act of shifting between the two languages causing an overly high mental load (Nawal, 2018). Therefore, to enhance learning, feedback should be carefully designed, considering both its content and delivery method, to limit the cognitive load placed on learners (Fyffe et al. 2015). However, it is still not clear what the best methods of feedback are in L2 writing tasks, specifically, to produce the best results from students by providing the maximum amount of informative feedback while reducing the cognitive load as much as possible, especially in terms of CAF measures, which have been understudied as a means of feedback.

Research Questions

Based on the discussion above, it is possible that automated feedback on a few specific points may help students to improve their writing ability. However, it is not yet known which pieces of feedback students will be able to uptake, which will lead to improvements in human rater scores, and how useful students find feedback about Complexity, Accuracy and Fluency (CAF) measures, in particular number of words (W), supporting detail markers (SDM), vocabulary variety (CTTR), and sentence length (MLS). Therefore, this study aims to answer the following research questions:

1. Did students show discernible changes in their human-rated writing scores or CAF measures (i.e., W, SDM, CTTR, MLS) from their pre- to posttest scores after receiving automated feedback?
2. Was there any association between changes in particular CAF measures (i.e., W, SDM, CTTR, MLS) and changes in human-rated writing scores following the provision of automated feedback?
3. What were students' perceptions of the automated feedback system that provided feedback based on CAF measures?

Methods

Participants and Educational Context

A total of 124 L1 Japanese EFL learners at Tohoku University gave informed consent to participate in this study and provided all the applicable data. Tohoku University is a Japanese university with a high national ranking and international standing in science and engineering. Therefore, most students here tend to be studious, and most of the student body leans towards science majors. The learners in this study were largely CEFR B1 level with some B2 level students, as indicated by their TOEFL ITP(R) scores, taken one week before the start of this study (463–637; $M = 517.7$; $SD = 30.2$). The participants came from various majors (agriculture: 32, dentistry: 5, literature: 16, medicine: 9, education: 2, science: 26, engineering: 27, pharmacy: 7) taking a required first year general education English as a Foreign Language class in academic reading and writing as designated by the Tohoku University curriculum. All the students had Japanese as their first language and had studied English for at least

six years before entering the university. The class took place in the second semester, so all participants had at least one semester experience with academic English including learning to write academic summaries in the first semester English class. Only the data for students who gave informed consent was used in the study, but all other students still participated in all the activities mentioned as it was part of their regular course.

Automated Feedback Tool

To create an automated feedback tool based on CAF measures that would hopefully not be too cognitively demanding, we first selected measures to students that would be: (1) previously shown to correlate to human rater scores, (2) relevant and reasonably intuitive to students, and (3) easy enough to replicate with simple web-based programming languages. We also decided not to provide feedback for too many measures, as the more feedback given would increase the cognitive load placed on the students, who would have to interpret the scores and then amend their writing based on the measures provided to them. Therefore, we attempted to create a simple HTML/JavaScript based program that would provide one representative measure of fluency and one of each major type of complexity (i.e., lexical and syntactic). Accuracy measures were not used, as calculated measures do not show as strong of a correlation to L2 proficiency and have thus not received as much attention in the CAF studies of writing (e.g., Wolfe-Quintero, 1998). Furthermore, as students were asked to use specific discourse markers in line with their textbook, one representative measure of designated discourse marker usage was also added to the tool, for a total of four CAF measures on which feedback was provided. The code for the tool can be found at https://github.com/springuistics/Paragraph_Feedback, and an operating example can be found at https://springsenglish.online/writing/paragraph.html.

The fluency measure that we selected for use in the tool was the number of words (W), based on the fact that this is both the easiest measure to calculate, and also one of the most highly correlated measures to human rater scores on its own across a myriad of studies (e.g., Spring & Johnson, 2022; Kyle, 2016; Lu, 2011; Ortega, 2003; Wolfe-Quinterno et al., 1999). Furthermore, since the writing activities were timed, we felt it was reasonable to consider this a measure of fluency, following Lu (2010) and Wolfe-Quintero et al. (1998).

The lexical complexity measure that we selected for use in the tool to examine vocabulary variety was the corrected token-type ratio (CTTR), as calculated by Carroll (1964), Lu (2012), and Spring and Johnson (2022). This measure looks at the number of different words used and the diversity of the vocabulary. It was selected because it exhibited the highest amount of correlation to human rater scores in both Lu (2012) and Spring and Johnson (2022) and can be easily recreated with simple lemmatization without the need for syntactic parsing. It should be noted that Kyle et al. (2023) have more recently suggested that a version of the moving average token-type ratio (MATTR) set to 11 words rather than the 50 proposed by Johnson (1944) and used in Lu's (2012) lexical complexity analyzer has less cross-correlation with W than CTTR and is therefore a better pure measure of lexical complexity. However, we conducted this experiment before the release of Kyle et al. (2023), it was based on speaking test scores rather than writing, and the MATTR-11 showed correlation comparable to that of CTTR (i.e., Spring & Johnson, 2022), so we believe that the use of CTTR is still informative for the purposes of this paper, although we should potentially consider revisiting this in the future.

The syntactic complexity measure selected for use in the tool was the mean length of sentence (MLS). This measure was selected mostly since it can be easily reproduced in web-based applications (i.e., it does not require syntactic parsing) and shows a reasonable amount of correlation to human rater scoring (e.g., Kyle & Crossley, 2018; Lu, 2011; Spring, 2023). Though it should be noted that Kyle and Crossley (2018) showed that fine-grained measures of syntactic complexity can be combined to

make a model much more predictive of L2 writer scores than large-grained measures, such measures require dependency tagging and would have to be combined into an integrated model that would likely be much more difficult for students to intuitively understand, whereas most of them can quickly look at their sentences and realize which have a smaller number of words than others.

Finally, the discourse marker measure selected for the tool was the simple count of the specific words and phrases presented in the students' textbook, called supporting detail markers (SDMs) in Spring (2023). This measure was selected mostly because of its immediate relevance to the students and their existing curriculum and the relative ease with which students can understand it. Furthermore, this measure is easily reproducible in web-based applications and was shown to be slightly correlated to human rater scores and have some modicum of importance to an integrated model that predicts human rater scores (Spring, 2023).

The tool provided feedback to the students on each of the four measures in both written and graph form. As all participants were English language learners, we felt that it was important to also give the students a visual representation of their feedback especially since one of the biggest problems students have with automated feedback tools is that the language level was too difficult (Taskiran et al., 2024). Four separate bar graphs were provided, each representing a different measure. Within each graph, students could compare three scores: a minimum acceptable score, their own score, and a 'good' score to aim for such as shown in Figure 1. Both the minimum acceptable score and the 'good' score were assigned based on previous data from students at the same university conducting a similar task (see Spring & Johnson, 2022; Spring, 2023), i.e., both were set at 2 standard deviations below and above, respectively, the median score from other data taken.

Classroom Procedure

Students participating in the study were in one of two classes taught by the authors, both native speakers of English. The authors taught from a unified curriculum and as such used the same textbook and teaching materials. The study consisted of 5 Classroom sessions of 90 minutes each. At the start of

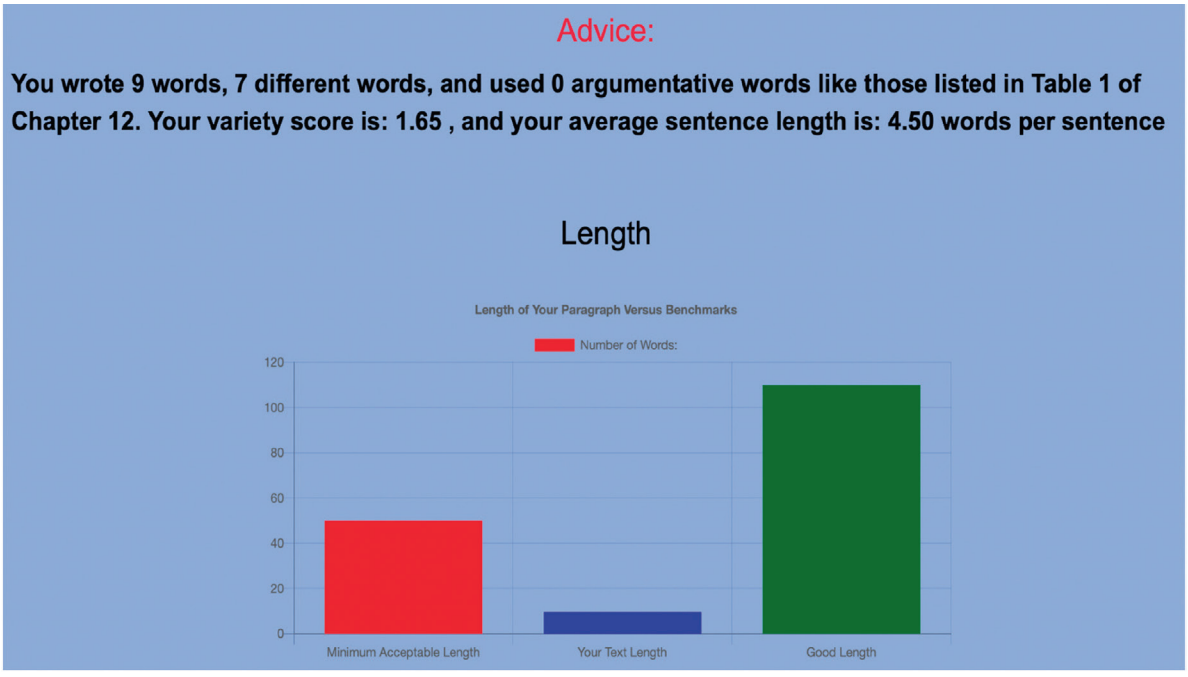

Figure 1 *Example of the Feedback on the Fluency Measure (Number of Words).*

each session the students were given 15 minutes to write an academic opinion paragraph responding to a prompt. Students were able to use any standard word processing document to type their paragraph. Each week the students then were taught about topics that would help them improve their paragraph writing through their textbook and the use of practice worksheets. The topics, in order, were paragraph structure, providing evidence, marking supporting details with discourse markers, writing with a variety of vocabulary and grammatical structures, and avoiding logical fallacies. After completing practice, and receiving training on the topic of the day, students then attempted to rewrite their original paragraph. After they felt happy with their paragraph they copied and pasted it into the automated feedback tool. This would not only give them their results but also allow them to check if they needed further rewrites. As we wanted all students to try and rewrite their paragraphs after learning a new concept, we felt that students should attempt to rewrite their paragraph before receiving a score and use the tool as a confirmation check afterwards. After checking the feedback tool students would then rewrite their paragraph again trying to improve their scores. Finally, students took part in peer review and gave each other advice on how to improve their automated feedback scores before submitting the final version at the end of the class.

Data Collection and Analysis

Before starting the study, all students wrote a paragraph in 15 minutes about a specified prompt as a pretest. After the 5 classroom sessions the students repeated the paragraph writing activity with a different prompt as the posttest. The essays were randomly mixed and given to 4 human raters. The raters were not informed that these essays were pre and post data in order that their perceptions on the essays were not influenced. The CAF measures (W, SDM, CTTR, MLS) were measured for pre- and posttest data using the tool provided by Spring and Johnson (2022), which is a combination of Lu's (2010; 2012) previous tools and includes some measures from Kyle (2016) but that uses SpaCy as the LLM to first parse the text. This code and an explanation can be found at https://github.com/mwjohnson/autograder.

After completing the final test, the participants were asked to answer a survey about their experience using the automated feedback tool. The survey items aimed to gather information on the students' perception of the cognitive load needed to use the tool. The cognitive load of students was measured subjectively by asking them to self-report their experience of cognitive load on a Likert-type scale. Part of the reason for this decision was that assessing cognitive load is generally seen as challenging (de Jong, 2010; Martin, 2014), with some researchers even questioning whether different measurements are needed for the three different types of cognitive load (Kalyuga, 2011). Methods of measurement can be split into three main types: subjective, behavioral, and physiological (Skulmowski & Rey, 2017). The method of measuring cognitive load that is most used is participants self-report using subjective surveys (Nawal, 2018). These surveys tend to use rating scales such as Likert and have been found to not only be easy to implement in the classroom but also reliable and sensitive (Pass et al., 1994). Behavioral measures can include analysis of reaction times or the tracking of eye movements while physiological measures look at heart rate, pupil dilation and brain imaging (Skulmowski & Rey, 2017). Therefore, with all limitations considered, in a normal classroom environment subjective measurement of cognitive load, such as the survey used, is the most realistic. The language for the survey was adapted from existing survey research (Zu. et al., 2021) and focused on measuring the students' extraneous and germane cognitive load. The participants rated their agreement with statements about their mental effort on a 5-point scale (1 – I don't think so, 5 – I think so). There was also one open-ended question at the end of the survey for the students to self-report their opinions on the feedback tool.

To assess the first research question, we used simple pre-post statistical testing of the human rater scores and the four CAF measures for which students were provided feedback, i.e., W, CTTR, SDM,

and MLS. The pre- and posttest data were checked for normality using a Shapiro-Wilk test and any data sets that were normally distributed were tested using *t*-tests with Cohen's *d* provided as a measure of effect size, and those that failed were tested using Wilcoxon signed-rank tests with *rs* used for effect size. Effect sizes were interpreted according to Plonsky and Oswald (2014), using the cut-offs provided by Spring (2022a).

To assess the second research question, we created a regression model with human-rater delta scores (i.e., posttest minus pretest) as the dependent variable and the delta scores of each of the four CAF measures as predictor variables. Initial correlation between change in human rater scores and change in CAF measures was provided via Pearson correlation tests, and then the regression model was created following Mizumoto (2023), using dominance analysis to calculate the relative importance of each predictor and random forests to determine whether each was significant to the model.

Finally, the open-ended results of the survey were analyzed qualitatively and the answers to the Likert-scale and selection questions were simply analyzed descriptively. These results were then considered in combination with the statistical approaches to provide triangulation and better interpret the data holistically.

Results

The descriptive data for pre- and posttests are shown in Table 1, and the results of the pre-posttest analyses are shown in Table 2. They reveal that overall, the students tended to show improvement in their human rater scores, with a medium effect size. Furthermore, students could improve their fluency, as measured by the number of words (W), syntactic complexity, as measured by mean length of sentence (MLS), and usage of discourse markers, as measured by the amount used from lists provided in their textbooks (SDM), all with large effect sizes. However, students were not able to improve their overall lexical complexity, as measured by CTTR.

Table 1 *Descriptive Statistics for Pre- and Posttest Scores*

Measure	Pretest Range, *M, SD*	Posttest Range, *M, SD*	Delta Range, *M, SD*
HR	4–20; 11.74, 3.06	8–20; 13.44, 2.31	–6–12; 1.70, 3.24
W	28–226; 84.7, 31.1	63–210; 116.1, 27.1	–52–100; 31.4, 28.3
SDM	0–8; 2.40, 1.47	1–12; 4.90, 2.19	–4–9; 2.50, 2.41
CTTR	3.04–5.34; 4.16, .45	2.41–5.25; 4.22, .45	–1.02–1.84; .06, .53
MLS	7–27; 13.63, 3.23	11–40; 18.08, 4.71	–12–28; 4.45, 5.29

Table 2 *Significance Testing Between Pre- and Posttest Scores*

Measure	Significance Test
Human-raters (HR)	$Z = 5.22$, $p < .01$, $rs = .50$**
Number of words (W)	$Z = 8.41$, $p < .01$, $rs = .76$***
Discourse markers (SDM)	$t = 11.55$, $p < .01$, $d = 1.04$***
Corrected token-type ratio (CTTR)	$t = 1.35$, $p = .18$, $d = .12$
Mean length of sentence (MLS)	$Z = 8.02$, $p < .01$, $rs = .72$***

Note: *Z* values indicate the results of Wilcoxon signed-rank tests, ***large effect size, **medium effect size.

The results of the simple correlation analyses between pre- and posttest human rater scores and each of the four CAF measures' delta scores is shown in Table 3. These initial results suggest that students who became able to write more words, use more discourse markers, and increase their vocabulary variety, were more likely to improve their human rater score as well, although the same cannot be said for improving on the mean length of sentence.

The regression analysis model was able to significantly predict change in human-rater scores, explaining 56% of the variance; $F = 37.808$, $p < .001$, $R2 = .56$. The results of this analysis are shown in Table 4 and Figure 2 and suggest that increases in total number of words written (W) and vocabulary variety (CTTR) were the most indicative of students whose human rater scores also improved. The improvement of mean length of sentence (MLS) was found to also be slightly important, but far less so than W or CTTR. Finally, increasing the number of discourse markers (SDM) was a poor predictor of improvement on human rated scores.

Table 3 *Correlation Between Change in CAF Measures and Change in Human Rater Scores*

	W	SDM	CTTR	MLS
HR	.72**	.20*	.35**	.07

Note: *$p < .05$, **$p < .01$.

Table 4 *Regression Analysis for Change in Human Rater Score Predictor Data*

Measure	B	Beta	*t*	*p*	RI	RF
W	.082	.713	10.51	.0001	82%	Confirmed
SDM	–.128	–.095	–1.44	.153	4%	Rejected
CTTR	1.127	.184	2.92	.004	14%	Confirmed
MLS	–.012	–.019	–.31	.758	0%	Tentative

Note: RI = relative importance, RF = random forest.

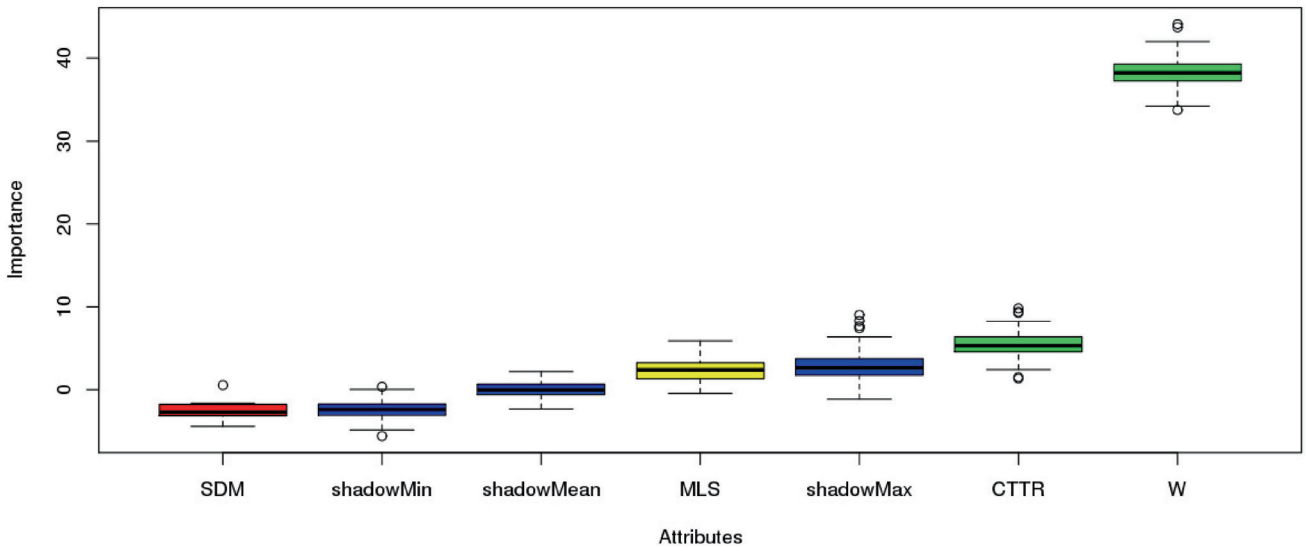

Figure 2 *Graphical Representation of Importance and Significance of Predictor Variables.*

The results of the posttest student survey can be seen in Tables 5 and 6. The first questions on the survey focused on assessing the students extraneous cognitive load by asking about the presentation of the feedback with questions concerning ambiguity and confusion. Students were first asked if they felt there was distracting (i.e. unnecessary) information in the feedback provided. Based on a 1–5 scale, there was very little information deemed distracting (M =1.43) and when asked to choose any of the four CAF measure charts that were distracting the results were 3% or lower. However, there were some instances where students did not understand the feedback generated and found it confusing (M = 2.32). While most of the CAF feedback graphs had low reported rates of confusion (i.e., Words Written (W) = 2%, Vocabulary Variety (CTTR) = 13%, Mean Length of Sentence = 6%), responses showed that 48 students (39%) found the chart for supporting detail markers (SDMs) confusing. Student comments in the final open-ended question suggest that students were unsure how this measure was calculated rather than a problem with the chart design itself with a student stating, "I don't know what formulas are used to measure the amount of evidence and lexical diversity." and another student writing "It would be nice if there are more detailed explanation about Amount of Evidence and Vocabulary Variety" referencing the SDM and CTTR charts.

The next questions in the survey were intended to measure the students' germane cognitive load by asking about their opinion on the mental demand placed on them by the feedback. Students, in general, did not find it hard to identify relevant information (M = 2.01) and expressed that it did not take a lot of mental effort to find the relevant information (M = 1.85). Also, when survey answers were

Table 5 *Descriptive Data for Likert-Scale Survey Questions*

Statements	No answer	1 "I don't think so"	2	3	4	5 "I think so"	*M*
"There was a lot of confusing information in the charts."	0	40	40	16	20	8	2.32
"There was a lot of distracting information in the charts."	1	89	21	8	3	2	1.43
"It was hard to identify relevant information in my feedback."	0	58	30	21	7	8	2.01
"It took a lot of mental effort to find relevant information in my feedback."	0	64	28	21	9	2	1.85

Table 6 *Descriptive Data for Survey Questions Asking Students Ratings of Each Chart*

Questions	W	CTTR	SDM	MLS
Which (if any) charts were confusing?	2% (3)	13% (16)	39% (48)	6% (8)
Which (if any) were distracting?	3% (4)	1% (1)	3% (4)	2% (3)
Which (if any) of the charts were helpful?	47% (52)	81% (90)	52% (58)	51% (57)

Table 7 *Student Responses to the Request for Comments on the Automated Feedback Tool*

Comment Type	No. of Comments
Comments thanking the tool developers	2
Comments praising the tool	10
Suggestions for improvement	6
Complaints	2

cross-referenced with pre- and posttest data, there was no correlation between negative feelings (i.e., self-reported cognitive load) and improvement in any measure or in human-rater scores. Therefore, the automated feedback was likely not overly demanding.

Students were then asked to select which of the four charts they found helpful and were able to choose multiple charts at the same time. All charts were selected by roughly half the students or more, but the most popular chart was vocabulary variety (CTTR) (81%). Students also mentioned its usefulness in their responses to the final open-ended question with one student stating, "Especially the variety of vocabulary matrix helped me a lot." However, in comparing which graphs students found helpful versus which measures were found to correlate to improvement in human-rater scores, students might not have a good idea of what is most important or helpful for them. Specifically, the order of importance for improving human rater scores was words written, vocabulary variety, SDM/MLS, but the order of graphs that the students rated as most helpful was vocabulary variety, SDM/MLS, words written. Therefore, the information might not be distracting, but they might need more guidance.

The final question in the survey was open-ended and asked students for any comments they had about the feedback tool and its use in class. Out of the 124 students who answered the survey, only 15 responded to this question but their comments expressed 20 different 'ideas' in total as can be seen in Table 7.

In general, the students' positive comments expressed that they found the tool helpful in improving their writing and they liked quickly knowing what areas they needed to focus on to improve their paragraphs. Both suggestions for improvements and complaints mainly focused on the desire for more information. Students wanted more information on how their writing was being graded, they requested specific areas of concern within their writing be highlighted by the tool and they also wanted concrete examples on how to improve.

Discussion

The results of this paper, taken together, offer potential insights by providing answers to the research questions and reveal several areas that should perhaps be addressed in the future.

Student Improvement on Human-rated Writing Scores and CAF Measures

To answer the first research question, it is clear that the treatment had a large impact on students' ability to write an opinion paragraph, as indicated by the significant differences in pre- and posttests with a medium effect size for human rater scores and large effects sizes for improvement in number of words written, number of discourse markers used, and average sentence length. While it is somewhat unclear what percentage of this increase was due to the use of the automated feedback tool, specifically, as students also received teacher instruction and conducted practice sessions and

peer review (i.e., Lu et al., 2021; Ngyuen & Tran, 2023), the fact that students were largely favorable towards the automated feedback and found it helpful indicates that it had at least some impact on their writing. Therefore, we believe that providing some objective measures of CAF measure feedback to students is likely enough to help them reconsider their writing and work to improve it; both in terms of specific CAF measures and in terms of general ability, as measured by human-raters. However, it should be noted that the only CAF measure that students could not improve upon, on average, was vocabulary variety. This is potentially because while students can perhaps make conscientious choices to use more of the specific target discourse markers in their books, more words in general, or more grammatical structures that extend sentence length, using a greater variety of vocabulary requires not only the awareness that the student should expand their vocabulary variety, but also actual knowledge of a range of words. Therefore, it might be that providing measures of vocabulary variety is not as important for students, or that it is only relevant at certain proficiency levels, when students have enough background vocabulary knowledge to be able to use more variety when told to do so. Another potential explanation could be that since vocabulary variety was one of the last topics that students were taught, students might have simply needed more time to practice this concept before being able to use it readily in their writing.

Association Between Changes in Particular CAF Measures and Human-rated Scores

With regards to the second research question, increases in the number of words written and in vocabulary variety were the two largest predictors of increases in human rater scores, with the results of mean length of sentence and discourse marker usage being inconclusive, i.e., the former was tentative in the regression model but showed little direct correlation, the latter showed higher direct correlation but was rejected in the regression model. Improvement in the number of words being correlated with increases in human rating is somewhat expected because this was one area that increased from pre- to posttests quite dramatically, but also because the number of words written is often a fairly good predictor of human-rater scores in general (e.g., Kyle, 2016; Lu, 2012; Spring & Johnson, 2022). However, the fact that vocabulary variety (i.e., CTTR) was the next most predictive measure of increases in human rater scores is surprising, partly because on average, students did not show much improvement in this area. This means that while it was difficult for most students to improve their vocabulary variety, those that did were able to improve their human-rater scores more than those who did not, indicating the importance of increasing this aspect of writing. Increases in mean length of sentence being only tentatively related to improvement in human rater scores might be explained in part by the fact that MLS is generally less predictive of writing scores than other metrics (e.g., Kyle & Crossley, 2018; Lu, 2012). Furthermore, it might be more important to provide new or more precise measures of syntactic complexity, although it might be challenging to create a predictive measure that is also intuitive to learners; although the verb-argument construction measures produced by Kyle (2016) can be combined to be very predictive of human rater scores (Kyle & Crossley, 2018; Lu et al., 2021), many are not easily understood by users outside of the field. Finally, the reason that an increase in the number of discourse markers (SDM) did not better correlate with human rater scores could simply be due to the markers selected. As pointed out by Spring (2023), there is a need for a better list of supporting detail discourse markers that show high correlation with human raters in general.

Students' Perceptions Towards the Automated CAF Measure Feedback

To answer the third research question, we examined the results of the student survey. We were concerned that the design of the automated feedback tool may cause the students to experience cognitive overload and be unable to process the feedback they received. In particular, we hypothesized that due to having to process multiple CAF measures of feedback while completing

the already cognitively demanding task of writing in the L2 (Nawal, 2018) the students may become overwhelmed and give up. However, our results showed that this was not the case. Most students reported very low overall germane and extraneous cognitive load. Furthermore, the lack of association between reported cognitive load and human-rater scores or improvement in any measure showed that the feedback method did not affect the students' ability to complete the task. The lack of predicted cognitive overload could be explained by the way the teachers introduced the tool to the students. Earlier in the semester students tended to be directed to focus mainly on the sentence length (MLS) and number of words written (W) graphs as supporting detail markers and vocabulary variety had not been studied yet. After each of those topics were addressed in class students would then actively use that part of the tool. This introduction to the tool in small steps may have helped students from feeling overwhelmed. However, while students did not appear to be overloaded, we did receive useful feedback in that some students were confused by the SDM measures, so either this measure should be made clearer in the future, or teachers should explain more clearly what this measure is. Furthermore, it was interesting that students found the vocabulary variety chart the most helpful, even though it was the area in which students improved the least. One reason for this could be that the chart helped students recognize that they needed to increase their vocabulary variety, and this became a good starting point, although improving on this measure might have simply been too difficult. As discussed above, one reason for this difficulty, as compared to other measures, is that it requires a larger vocabulary to implement, which is not necessarily the case for the other CAF measures. We were also concerned that the feedback might not have been specific enough, but the results of the survey, especially when considered with the results of the pre- and posttest data, suggest that this was not the case.

Pedagogical Implications

Based on the results and discussion above, we believe that providing students automated feedback based on CAF measures of their writing can be quite helpful both in terms of improving their individual CAF measures and their overall writing skills. However, it should be noted that there is some mismatch in the category of vocabulary variety. Specifically, though we found that improving vocabulary variety was very important for improving human rater scores, and that students said the feedback on vocabulary variety was the most helpful of all the measures, most students were not able to improve meaningfully on it from the pre- to posttest. This suggests that while providing practice and guidance is enough to help students improve on most CAF measures, vocabulary variety is more problematic and should therefore receive more attention in the classroom. It is possible that it needs to be taught as one of the first concepts in the semester to give students more time to practice it. Alternatively, teachers should consider teaching more synonym vocabulary that is very specific to the writing assignments they are giving to students to help them expand their variety when writing. It is also possible that further adjustments could or should be made to the tool in the future to increase its usefulness. For example, overused words could be highlighted and hyperlinked to a thesaurus page or other useful tools.

Furthermore, the results of this study suggest that while the number of words students write and their degree of vocabulary variety does impact their general writing ability, increasing the mean length of sentence and number of discourse markers might not be as important as previously believed. While further research is still needed in these areas, the current study suggests that teachers should take more steps to help improve the number of words in their students' written responses and the amount of vocabulary variety that they use. Alternatively, teachers may focus on improving the number of words written in lower-level students' writing and then focus on improving vocabulary variety and other aspects of higher-level learners who have arguably acquired enough L2 ability that they might be able to improve such areas with minimal feedback and instruction.

Limitations and Future Research

While this study has presented some interesting results that have practical pedagogical implications for EFL teaching and learning, it should be noted that there are also limitations to the study and areas that should be addressed in the future. First, it should be noted here that the CAF measures provided to students in this study were not at all comprehensive, nor were they designed to be. It does provide evidence that simply helping students to see which areas of CAF they should improve upon and providing them with goals helps them to improve their writing in terms of both CAF and human-rated scores. However, many of the measures provided by this tool were large-grained, and more specific advice could have been provided to help learners improve upon the metrics provided. As pointed out by Kyle (2016), fine-grained CAF measures are also an important consideration, and as Rastgou (2024) points out, both the scope and consistency of the feedback should be taken into consideration. Therefore, future studies should look at how automated feedback on other CAF measures impact learners' writing ability. Such studies should consider how understandable the measure is to the stakeholders, i.e., teachers and students, and how important that truly is. Furthermore, they should also consider how direct the automated feedback should be. For example, it is possible to adjust the tools to easily indicate over-repeated words or excessively short sentences to students. However, providing too much feedback might also hinder students' improvement, as they might become over-reliant on the tool, as is sometimes reported with grammar correcting tools (e.g., Gao & Ma, 2022; Shang, 2022; Shi & Aryadoust, 2024). Therefore, future studies should also consider how direct and how much feedback should be given for maximum improvement without creating dependence on the tools. This seems especially relevant in the age of large language models such as Chat GPT that can potentially automatically correct a variety of problems in students' writing without having them be very involved with the writing process at all.

Furthermore, there were also some limitations with the survey used to gain students' perceptions of the CAF measure automated feedback. In particular, the timing of the survey at the end of the writing course meant that the survey was administered to students just after studying about vocabulary variety. This led to doubts amongst the researchers that the vocabulary variety CAF measure was deemed as 'most helpful' because it had been studied most recently. However, due to classroom constraints it was not possible to complete the survey at any later date. Another limitation with the survey was the types of cognitive load questions. While questions targeting extraneous and germane load were included there were no questions that focused on intrinsic cognitive load (i.e. how complex the students found the task itself). Ideally, future research on cognitive load and automated feedback should seek to determine all types of cognitive load on students. This will become especially important if future studies work with more complex CAF measures that might potentially be less intuitive to teachers and students.

Conclusion

The results of this study show that after five treatment sessions with the automated feedback tool, most students were able to improve the number of words they were able to write, the number of supporting detail discourse markers they used, and the average length of their sentences, but that most were not able to improve their vocabulary variety. Furthermore, students' post-treatment writings did receive higher marks from human raters than their pre-treatment writings. Additionally, we found that increases in both the number of words written, and vocabulary variety were significantly predictive of increases in human rater scores, and that increases in sentence length were only tentatively predictive, whereas increased usage of discourse markers was not. Finally, the participants were generally positive towards the CAF measure based automated feedback tool, but work could be done to make these more intuitive and future works should also more carefully

consider which CAF measures to use in automated feedback tools to maximize student improvement in short-term treatments. However, in general, we found that the tool created was quite successful in the classroom, and more work should be done to improve and expand it in conjunction with teaching and other technologies in the future.

References

Akin, I. & MurrellJones, M. (2018). Closing the gap in academic writing using the cognitive load theory. *Literacy Information and Computer Education Journal,* 9, 2833–2841. https://doi.org/10.20533/licej.2040.2589.2018.0373

Benali, A. (2021). The impact of using automated writing feedback in ESL/EFL classroom contexts. *English Language Teaching, 14*(12), 189–195. https://doi.org/10.5539/elt.v14n12p189

Carless, D. (2020). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education, 23*(2), 143–153. https://doi.org/10.1177/1469787420945845

Carroll, J. B. (1964). *Language and thought.* Prentice-Hall.

Cen, Y., & Zheng, Y. (2024). The motivational aspect of feedback: A meta-analysis on the effect of different feedback practices on L2 learners' writing motivation. *Assessing Writing*, *59*, 100802. https://doi.org/10.1016/j.asw.2023.100802

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*, 14–27. https://doi.org/10.3758/s13428-018-1142-4

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instr Sci, 38*, 105–134. https://doi.org/10.1007/s11251-009-9110-0

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*, 1–53. http://dx.doi.org/10.1007/s10639-023-12402-3

Eguchi, M., & Kyle, K. (2023). Span identification of epistemic stance-taking in academic written English. In *Proceedings of The 18th Workshop on Innovative Use of NLP for Building Educational Applications*. https://doi.org/10.48550/arXiv.2306.02038

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System*, *36*, 353–371. https://doi.org/10.1016/j.system.2008.02.001

Fan, N., & Ma, Y. (2022). The effects of Automated Writing Evaluation (AWE) feedback on students' English writing quality: A systematic literature review. *Language Teaching Research Quarterly*, *28*, 53–73. https://doi.org/10.32038/ltrq.2022.28.03

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2015). When feedback is cognitively-demanding: the importance of working memory capacity. *Instr Sci, 43*, 73–91. https://doi.org/10.1007/s11251-014-9323-8

Gao, J., & Ma, S. (2022). Instructor feedback on free writing and automated corrective feedback in drills: Intensity and efficacy. *Language Teaching Research*, 26(5), 986–1009. https://doi.org/10.1177%2F1362168820915337

Guo, Q., Feng, R., & Hua, Y. (2021). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*, https://doi.org/10.1080/09588221.2021.1879161

Jiang, J., Bi, P., Liu, H. (2019). Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing, 46*, 1–13. https://doi.org/10.1016/j.jslw.2019.100666

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs, 56*, 1–15.

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37*, 39–56. https://doi.org/10.1016/j.asw.2018.03.002

Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2023). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition*, 1–22. https://doi.org/10.1017/S0272263123000402

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal, 102*(2), 333–349. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*, 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Unpublished doctoral dissertation, Georgia State University].

Li, M. (2021). Researching and teaching second language writing in the digital age. Springer Nature. https://doi.org/10.1007/978-3-030-87710-1_7

Lu, X., Ren, W., & Xie, Y. (2021). The effects of online feedback on ESL/EFL writing: A meta-analysis. *The Asia-Pacific Education Researcher, 30*, 643–653. https://doi.org/10.1007/s40299-021-00594-6

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2),190–208. https://doi.org/10.1111/j.1540-4781.2011.01232.x

Martin, S. (2014). Measuring cognitive load and cognition: Metrics for technology-enhanced learning. *Educational Research and Evaluation, 20*(7–8), 592–621. https://doi.org/10.1080/13803611.2014.997140

McNamara, D. S., Graesser, A. C., McCarthy, P. M., Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Mizumoto, A. (2023). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning, 73*(1), 161–196. https://doi.org/10.1111/lang.12518

Nawal, A. F. (2018). Cognitive load theory in the context of second language academic writing, *Higher Education Pedagogies*, *3*(1), 385–402. https://doi.org/10.1080/23752696.2018.1513812

Nguyen, T. T., & Tran, A. V. (2023). Using Write&Improve to improve EFL learners' writing skills. *CALL-EJ, 24*(1), 198–204. https://old.callej.org/journal/24-1/Nguyen-Tran2023.pdf

Niu, R., Shan, P., You, X. (2021) Complementation of multiple sources of feedback in EFL learners' writing. *Assessing Writing, 49*, https://doi.org/10.1016/j.asw.2021.100549

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Paas, F. G., Van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, *79*(1), 419–430. https://doi.org/10.2466/pms.1994.79.1.419

Plonsky, L., & Oswald, F. L., (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878–912. https://doi.org/10.1111/lang.12079

Rastgou, A. (2022). How feedback conditions broaden or constrain knowledge and perceptions about improvement in L2 writing: A 12-week exploratory study. *Assessing Writing, 53*, 100633. https://doi.org/10.1016/j.asw.2022.100633

Rastgou, A. (2024). Prospective methodological considerations in L2 written feedback research. *Feedback Research in Second Language, 1*, 175–196. https://doi.org/10.32038/frsl.2023.01.10

Shang, H. (2022). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments, 30*(1), 4–16. https://doi.org/10.1080/10494820.2019.1629601

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning, 64*(1), 103–131. https://doi.org/10.1111/lang.12029

Skulmowski, A. & Rey, G. D. (2017). Measuring cognitive load in embodied learning settings. *Frontiers in Psychology*, *8*, Article 1191, https://www.doi.org/10.3389/fpsyg.2017.01191

Spring, R. (2022). Free, online multilingual statistics for linguistics and language education researchers. *Center for Culture and Language Education, Tohoku University 2021 Nenpo, 8*, 32–38. https://doi.org/10.13140/RG.2.2.12037.63202

Spring, R. (2023). Transformations of number of words and phrases signaling supporting details: Potential variables for automated rating. *Language Education & Technology, 60*, 1–24. https://doi.org/10.24539/let.60.0_1

Spring, R., & Johnson, M. W. (2022). The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK, and SpaCy tools. *System, 106*, 770–786. https://doi.org/10.1016/j.system.2022.102770

Suzuki, W. (2012). Written languaging, direct correction, and second language writing revision. *Language Learning, 62*(4), 1110–1133. https://doi.org/10.1111/j.1467-9922.2012.00720.x

Suzuki, W., Nassaji, H., & Sato, K. (2019). The effects of feedback explicitness and type of target structure on accuracy in revision and new pieces of writing. *System*, *81*, 135–145. https://doi.org/10.1016/j.system.2018.12.017

Sweller, J., Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Taskiran, A., Yazici, M., & Erdem Aydin, I. (2024). Contribution of automated feedback to the English writing competence of distance foreign language learners. *E-Learning and Digital Media*, *21*(1), 24–41. https://doi.org/10.1177/20427530221139579

Tomen, M. (2022). Automated essay scoring feedback in foreign language writing: Does it coincide with instructor feedback? *Interdisciplinary Language Studies, 4*(4), 53–62. https://doi.org/10.1016/j.asw.2014.03.006

Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics, 38*(1), 90–111. https://doi.org/10.1093/applin/amv002

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). Second language development in writing: Measures of fluency, accuracy, and complexity. University of Hawaii Press. https://doi.org/10.1017/s0272263101263050

Yang, H., Gao, C., & Shen, H. Z. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. *Education and Information Technologies*, *29*(4), 3837–3858. https://doi.org/10.1007/s10639-023-11991-3

Zhang, J., & Zhang, L. J. (2022). The effect of feedback on metacognitive strategy use in EFL writing. *Computer Assisted Language Learning*, https://doi.org/10.1080/09588221.2022.2069822

Zu, T., Munsell, J., & Rebello, N. S. (2021). Subjective measure of cognitive load depends on participants' content knowledge level. *Frontiers in Education*, *6*, 647097. https://www.doi.org/10.3389/feduc.2021.647097